# PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2000-148793
(43)Date of publication of application : 30.05.2000

(51)Int.Cl. G06F 17/30

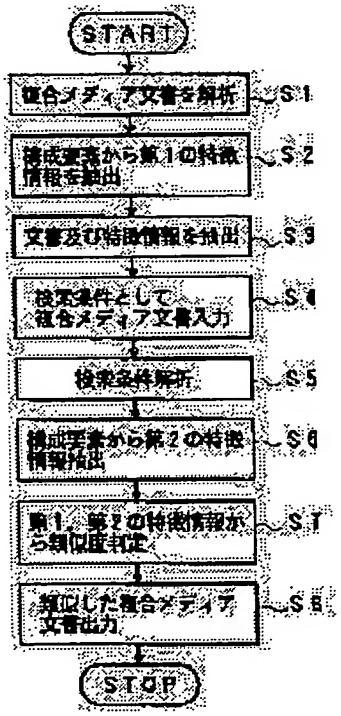
(21)Application number : 11-055950
(22)Date of filing : 03.03.1999

(71)Applicant : NIPPON TELEGR & TELEPH CORP <NTT>
(72)Inventor : HASEGAWA TOMOHIRO
UMEDA MASAYOSHI
TANIGUCHI NOBURO
YAMAMURO MASASHI

(30)Priority
Priority number : 10258763 Priority date : 11.09.1998 Priority country : JP

## (54) METHOD AND DEVICE FOR SIMILAR RETRIEVAL OF COMPOSITE MEDIA DOCUMENT AND STORAGE MEDIUM STORED WITH SIMILAR RETRIEVAL PROGRAM FOR COMPOSITE MEDIA DOCUMENT

(57)Abstract:
PROBLEM TO BE SOLVED: To perform retrieval by giving structure information to part of a retrieval key even if information regarding a document structure is not known by deciding the similarity of a composite media document according to feature information extracted from a constituent element of the composite media document and feature information extracted from a constituent element of a retrieval condition.
SOLUTION: The syntax of the given composite multimedia document is analyzed (S1) and 1st feature information is extracted from a constituent element of the composite media document (S2). The document and extracted feature information are stored (S3) and the composite media document is inputted as a retrieval condition (S4). The syntax of the retrieval condition is analyzed (S5) and 2nd feature information is extracted from a constituent element of the retrieval condition obtained as a result of the analysis (S6). Then the similarity between two composite media documents is decided (S7) according to the 1st feature information and 2nd feature information which are stored and similar composite media documents are outputted (S8).



* NOTICES *



1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

---

## CLAIMS

---

[Claim(s)]
[Claim 1]In a similar retrieval method of a compound media document which is a structured document which comprises speech information which adds to text information, picture information, and voice data that is data of human being's voice, and contains composition data containing CD and a record and music data, The 1st characteristic information is extracted from a component of said compound media document which analyzes the syntax of said given compound media document, and is obtained as a result of being analyzed, Accumulate said compound media document and said extracted characteristic information, and a compound media document is inputted as a search condition, The 2nd characteristic information is extracted from a component of said search condition which analyzes the syntax of said inputted search condition, and is acquired as a result of being analyzed, A similar retrieval method of a compound media document which judges similarity of two compound media documents based on said 1st characteristic information accumulated and said 2nd characteristic information, and is characterized by outputting a similar compound media document.
[Claim 2]A similar retrieval method of the compound media document according to claim 1 which inputs a document which a user illustrated as a search key at the time of search, extracts said 2nd characteristic information from an illustrated document, and calculates similarity between documents by said 2nd extracted characteristic information and said 1st characteristic information.
[Claim 3]A similar retrieval method of a compound media document given in claims 1 and 2 which set up an evaluation value based on a similarity decision result for every component with information on media and structure information characterized by comprising the following as whole compound media document similarity.
Information and structure information on media containing a text, a picture, and a sound which

constitute said document illustrated as said search key when calculating similarity of said compound media document.
A text, a picture, a sound which constitute said document accumulated.

[Claim 4]As a similarity decision result for said every component, text information of said illustrated document, A similarity decision result of text information of said accumulated document, and said illustrated picture information of a document, A similarity decision result with picture information of an accumulated this document, and speech information of a this illustrated document, A similar retrieval method of the compound media document according to claim 3 using a similarity decision result of a similarity decision result with speech information of an accumulated this document, structure information on a this illustrated document, and structure information on a this accumulated document.
[Claim 5]Text information included in said document when calculating similarity of said compound media document, A similar retrieval method of the compound media document according to claim 3 which makes what calculated similarity, applied a value of dignity to said similarity for every characteristic information of picture information, speech information, and structure information, and took linear combination similarity as said whole compound media document.
[Claim 6]When calculating similarity of said compound media document and two or more same media exist in the same document, About all the search keys for every media including text information, picture information, and speech information which are included in said illustrated document. A similar retrieval method of a compound media document given in claims 2 and 3 which calculate all the similarity to this search key in these media in an accumulated document and to which said similarity makes the highest thing similarity of representation to said search key.
[Claim 7]When calculating similarity of said compound media document, as said search key, Said tree with an order label which expressed each of structure information on said illustrated document, and structure information on said accumulated document as a tree with an order label (ordered labeled tree), and expressed said illustrated document, A similar retrieval method of the compound media document according to claim 1 of setting up similarity of structure information on a document by comparing shape with a tree with an order label expressing said accumulated document.
[Claim 8]When setting up similarity of said structure information, it is considered that said document structure is a tree, The number of times which performed editing operation including insertion of a node required in order to change into a tree showing said document accumulated from a tree showing said illustrated document, deletion of a node, and change of a node name, A similar retrieval method of the compound media document according to claim 7 of setting up

compilation distance computed from cost required to perform this editing operation as similarity of said document.

[Claim 9]Text information included in a document when calculating similarity of a compound media document, Similarity based on characteristic information of picture information, speech information, and structure information is calculated, A similar retrieval method of the compound media document according to claim 3 which makes similarity based on characteristic information of text information and picture information which perform the 1st-step selection based on a similarity calculation result based on characteristic information of structure information, and are contained in said document, or speech information similarity as the whole compound media document.

[Claim 10]Text information included in a document when calculating similarity of a compound media document, Similarity based on characteristic information of picture information, speech information, and structure information is calculated, A similar retrieval method of the compound media document according to claim 3 which makes similarity based on characteristic information of structure information which performs the 1st-step selection based on a similarity calculation result based on characteristic information of said text information, said picture information, or said speech information, and is included in said document similarity as the whole compound media document.

[Claim 11]A similar retrieval method of a compound media document given in claims 2 and 3 which set up similarity in a document level of these media when calculating similarity of a compound media document and two or more same media exist in a document illustrated as a search key.

[Claim 12]When setting up similarity in a document level of said media, about each search key of two or more said media in a document illustrated as a search key. About all the search keys for every media including text information, picture information, and speech information which are included in said illustrated document. All the similarity to this search key in these media in an accumulated document is calculated, A similar retrieval method of the compound media document according to claim 11 that said similarity makes the highest thing similarity of representation to said search key, calculates average value of similarity of said representation, and sets up similarity in a document level of said media.

[Claim 13]When setting up similarity in a document level of said media, about each search key of two or more said media in a document illustrated as a search key. About all the search keys for every media including text information, picture information, and speech information which are included in said illustrated document. All the similarity to this search key in these media in an accumulated document is calculated, A similar retrieval method of the compound media document according to claim 11 that said similarity makes the highest thing similarity of representation to said search key, and makes what has the highest similarity similarity in a

document level of said media among similarity of said representation.
[Claim 14]When performing similar retrieval of a compound media document, each of structure information on a document illustrated as a search key and structure information on an accumulated document is expressed as a tree with an order label (ordered labeled tree), Characteristic information of each media in each document is expressed as a tree with an order label with an attribute stored as an attribute of a node in said tree with an order label, A similar retrieval method of the compound media document according to claim 7 of setting up similarity of a compound media document by comparing an attribute and shape with a tree with an order label with an attribute expressing a tree with an order label with an attribute expressing said illustrated document, and said accumulated document.
[Claim 15]When performing similar retrieval of a compound media document, about a tree with an order label with an attribute expressing an accumulated document with a node which has characteristic information which is the attribute of each node of a tree with an order label with an attribute expressing said illustrated document, and similar characteristic information as an attribute. A similar retrieval method of the compound media document according to claim 14 of setting up similarity of a compound media document from a difference in structural physical relationship of said node.
[Claim 16]When setting up similarity of said structure information, it is considered that document structure is a tree with an order label, A similar retrieval method of the compound media document according to claim 7 of evaluating and mapping this characteristic information and setting up distance on said vector space as similarity of a document on multi dimensional vector space based on characteristic information about this tree with an order label.
[Claim 17]A similar retrieval method of the compound media document according to claim 16 which calculates similarity of a document by evaluating and using position information on a name and a node number of each node of said tree with an order label, or each node as characteristic information about said label tree with an order.
[Claim 18]If it is the frequency of occurrence of a concept which a descriptive content of a text expresses as said characteristic information if it is text information, or each word, and picture information, If it is hue of a picture, chroma saturation, luminosity, color arrangement, and speech information and is strength of a sound, a melody, and structure information, A similar retrieval method of the compound media document according to claim 1 which makes shape of a tree when a tree with an order label expresses document structure, a label name of a node, link information, etc. characteristic information from which it is extracted from a component of said compound media document.
[Claim 19]An evaluation value based on a similarity decision result for every component with information on media and structure information characterized by comprising the following, A similar retrieval method of the compound media document according to claim 1 of setting up as

whole compound media document similarity, carrying out ranking by arranging similarity of said accumulated document in a descending order, and judging similarity.
Information and structure information on media containing a text, a picture, and a sound which constitute said document illustrated considering similarity of an accumulated document to said illustrated document as said search key when judging said similarity.
A text, a picture, a sound which constitute said document accumulated.

[Claim 20]When setting up said similarity, for every component of said compound media document. Text information, picture information and speech information which set up similarity and are included in said document, And a similar retrieval method of the compound media document according to claim 3 which makes what calculated similarity, applied a value of dignity to said similarity for every characteristic information of structure information, and took linear combination similarity as said whole compound media document.
[Claim 21]When setting up similarity of said whole compound media document, as an evaluation value based on a similarity decision result for every component of said document The similarity of each component itself. Or a similar retrieval method of the compound media document according to claim 3 of using what multiplied dignity given to similarity of each component by user.
[Claim 22]A similarity retrieval device of a compound media document which is a structured document which adds to text information, picture information, and voice data that is data of human being's voice characterized by comprising the following, and comprises speech information containing composition data containing CD and a record and music data.
A compound media document input means which inputs a compound media document.
Said compound media document given by said compound media document input means, and a document analyzing means which analyzes the syntax of an inputted search condition.
A characteristic information extraction means to extract characteristic information from a component of a document obtained as a result of being analyzed by said document analyzing means.
An accumulation means which accumulates said characteristic information extracted by said compound media document and said characteristic information extraction means, A search condition input means which inputs a compound media document as a search condition, and characteristic information of said compound media document accumulated in said accumulation means, A document comparison means to judge similarity of two compound media documents based on characteristic information extracted by said characteristic information extraction means based on a result of having analyzed said inputted search condition by said document analyzing means, An output means which outputs a similar compound media document based on similarity judged by said document comparison means.

[Claim 23]A similarity retrieval device of the compound media document according to claim 22 characterized by comprising the following.
An input sentence document characteristic information extraction means to extract a compound media document illustrated as a search key including a means to input a document in which a user illustrated said search condition input means as a search key from said compound media document in which said characteristic information extraction means was given.
Input sentence document characteristic information by which said document comparison means was extracted from a document illustrated by said user by said input sentence document characteristic information extraction means including a search characteristic information extraction means to extract characteristic information of said search information.
A similarity calculation means to calculate similarity between compound media documents illustrated as said compound media document and said search key by search characteristic information extracted by said search characteristic information extraction means.

[Claim 24]A similarity retrieval device of a compound media document given in claims 21 and 22 characterized by comprising the following.
Text information, picture information, speech information, and structure information that said similarity calculation means constitutes said illustrated document from said search condition input means as said search key.
A similarity setting-out means to set up an evaluation value based on a similarity decision result for every component with text information, picture information, speech information, and structure information which constitute said compound media document accumulated in said accumulation means as similarity of the whole compound media document.

[Claim 25]A similarity decision result of text information of said document in which said similarity setting-out means is accumulated in text information and said accumulation means of said illustrated document as a similarity decision result for said every component, A similarity decision result of picture information of an illustrated this document, and picture information of a document accumulated in this accumulation means, A similarity retrieval device of the compound media document according to claim 24 using a similarity decision result of speech information of an illustrated this document, and speech information of a document accumulated in this accumulation means, and a similarity decision result of structure information on this illustration **** document, and structure information on a document accumulated in this accumulation means.
[Claim 26]Text information, picture information and speech information by which a front

similarity setting-out means is contained in said document, And a similarity retrieval device of the compound media document according to claim 24 including a linear combination calculating means which makes what calculated similarity, applied a value of dignity to this similarity for every characteristic information of structure information, and took linear combination similarity as said whole compound media document.

[Claim 27]When two or more same media exist in the same document, said document comparison means, About all the search keys for every media containing a text, a picture, and a sound which are contained in said illustrated document. A similarity retrieval device of a compound media document given in claims 22 and 23 containing a representation similarity determination means by which all the similarity to this search key in media in an accumulated document is calculated, and this similarity makes the highest thing similarity of representation to said search key.

[Claim 28]A similarity retrieval device of the compound media document according to claim 23 characterized by comprising the following.

Said tree with an order label which said similarity calculation means expressed each of structure information on said illustrated document, and structure information on an accumulated document as a tree with an order label (ordered labeled tree) as said search key, and expressed said illustrated document.

A tree shape comparison means with an order label to set up similarity of structure information on a document by comparing shape with a tree with an order label expressing said accumulated document.

[Claim 29]A similarity retrieval device of the compound media document according to claim 28 characterized by comprising the following.

The number of times which performed editing operation including insertion of a node required in order to change said tree shape comparison means with an order label into a tree showing said document accumulated from a tree which considers that said document structure is a tree and expresses said illustrated document when judging similarity of said structure information, deletion of a node, and change of a node name.

A compilation distance calculating means which sets up compilation distance computed from cost required to perform this editing operation as similarity of said document.

[Claim 30]A similarity retrieval device of the compound media document according to claim 24 characterized by comprising the following.

A means to calculate similarity based on characteristic information of text information, picture information and speech information, and structure information that said similarity calculation means is contained in a document.

A means to perform the 1st-step selection based on a similarity calculation result based on characteristic information of structure information.
A means which makes similarity based on characteristic information of text information and picture information which are contained in said document, or speech information similarity as the whole compound media document.

[Claim 31]A similarity retrieval device of the compound media document according to claim 24 characterized by comprising the following.
A means to calculate similarity based on characteristic information of text information, picture information and speech information, and structure information that said similarity calculation means is contained in a document.
A means to perform the 1st-step selection based on a similarity calculation result based on characteristic information of said text information, said picture information, or said speech information.
A means which makes similarity based on characteristic information of structure information included in said document similarity as the whole compound media document.

[Claim 32]A similarity retrieval device of a compound media document given in claims 23 and 24 which contain a document level similarity calculation means to set up similarity in a document level of these media when two or more same media exist in a document in which said similarity calculation means was illustrated as a search key.
[Claim 33]A similarity retrieval device of the compound media document according to claim 32 characterized by comprising the following.
Said document level similarity calculation means about each search key of two or more said media in a document illustrated as a search key. A means to calculate all the similarity to this search key in these media in an accumulated document about all the search keys for every media including text information, picture information, and speech information which are included in said illustrated document.
A means by which said similarity makes the highest thing similarity of representation to said search key.
A means to calculate average value of similarity of said representation.
A document level similarity setting-out means to set up similarity in a document level of said media.

[Claim 34]A similarity retrieval device of the compound media document according to claim 32 characterized by comprising the following.
Said document level similarity setting-out means about each search key of two or more said

media in a document illustrated as a search key. A means to calculate all the similarity to this search key in these media in an accumulated document about all the search keys for every media including text information, picture information, and speech information which are included in an illustrated this document.
A means by which said similarity makes the highest thing similarity of representation to said search key.
A means which makes what has the highest similarity similarity in a document level of said media among similarity of said representation.

[Claim 35]Said similarity calculation means expresses each of structure information on a document illustrated as a search key, and structure information on an accumulated document as a tree with an order label (ordered labeled tree), Characteristic information of each media in each document is expressed as a tree with an order label with an attribute stored as an attribute of a node in said tree with an order label, A similarity retrieval device of the compound media document according to claim 28 which contains a similar retrieval means to set up similarity of a compound statement document, by comparing an attribute and shape with a tree with an order label with an attribute expressing a tree with an order label with an attribute expressing said illustrated document, and said accumulated document.
[Claim 36]Said similar retrieval means about a tree with an order label with an attribute expressing an accumulated document with a node which has characteristic information which is the attribute of each node of a tree with an order label with an attribute expressing an illustrated document, and similar characteristic information as an attribute. A similarity retrieval device of the compound media document according to claim 35 containing a means to set up similarity of a compound media document from a difference in structural physical relationship of this node.
[Claim 37]A similarity retrieval device of the compound media document according to claim 30 characterized by comprising the following.
A means for said tree shape comparison means with an order label to consider that document structure is a tree with an order label, and to evaluate and map this characteristic information on multi dimensional vector space based on characteristic information about this tree with an order label.
A means to set up distance on said vector space as similarity of a document.

[Claim 38]A similarity retrieval device of the compound media document according to claim 37 which calculates similarity of a document by evaluating and using position information on a name and a node number of each node of said tree with an order label, or each node as characteristic information about said label tree with an order.

[Claim 39]If it is the frequency of occurrence of a concept which a descriptive content of a text expresses as said characteristic information if it is text information, or each word, and picture information, If it is hue of a picture, chroma saturation, luminosity, color arrangement, and speech information and is strength of a sound, a melody, and structure information, A similarity retrieval device of the compound media document according to claim 22 which makes shape of a tree when a tree with an order label expresses document structure, a label name of a node, link information, etc. characteristic information from which it is extracted from a component of said compound media document.
[Claim 40]A similarity retrieval device of the compound media document according to claim 22 characterized by comprising the following.
Information and structure information on media containing a text, a picture, and a sound which constitute said document in which said document comparison means was illustrated considering similarity of an accumulated document to said illustrated document as said search key.
A means to set up an evaluation value based on a similarity decision result for every component with information on media and structure information containing a text, a picture, and a sound which constitute said document accumulated as whole compound media document similarity.
A means to carry out ranking by arranging similarity of said accumulated document in a descending order, and to judge similarity.

[Claim 41]A similarity retrieval device of the compound media document according to claim 24 characterized by comprising the following.
A means by which said similarity setting-out means sets up similarity for every component of said compound media document.
A means to calculate similarity for every characteristic information of text information included in said document, picture information and speech information, and structure information.
A means which makes what applied a value of dignity to said similarity and took linear combination similarity as said whole compound media document.

[Claim 42]A similarity retrieval device of the compound media document according to claim 24 in which said similarity setting-out means contains a means to use the similarity of each component itself, or a thing which multiplied dignity given to similarity of each component by user, as an evaluation value based on a similarity decision result for every component of said document.
[Claim 43]A storage which added to text information, picture information, and voice data that is data of human being's voice characterized by comprising the following, and stored a similar

retrieval program of a compound media document which is a structured document which comprises speech information containing composition data containing CD and a record and music data.
Said given compound media document and a document analyzing process of analyzing the syntax of an inputted search condition.
A characteristic information extraction process of extracting characteristic information from a component of a document obtained as a result of being analyzed in said document analyzing process.
A storing process of storing in a memory measure said characteristic information extracted in said compound media document and said characteristic information extraction process.
A search condition input process of making a compound media document inputting as a search condition, Characteristic information of said compound media document accumulated in said memory measure, A document comparison process of judging similarity of two compound media documents from characteristic information extracted in said characteristic information extraction process based on a result of having analyzed said inputted search condition in said document analyzing process, An output process to which a similar compound media document is made to output based on similarity judged in said document comparison process.

[Claim 44]A storage which stored a similar retrieval program of the compound media document according to claim 43, comprising:
An input sentence document characteristic information extraction process of extracting characteristic information from said compound media document in which said characteristic information extraction process was given including a process of inputting a document in which a user illustrated said search condition input process as a search key.
Input sentence document characteristic information from which said document comparison process was extracted from a document illustrated by said user in said input sentence document characteristic information extraction process including a search characteristic information extraction process of extracting characteristic information of a compound media document illustrated as a search key.
A similarity calculation process of calculating similarity between compound media documents illustrated as said compound media document and said search key by search characteristic information extracted in said search characteristic information extraction process.

[Claim 45]A storage which stored a similar retrieval program of a compound media document given in claims 43 and 44, comprising:
Text information, picture information, speech information, and structure information that said

similarity calculation process constitutes a document illustrated as said search key from said search condition input process.
A similarity setting process which sets up an evaluation value based on a similarity decision result for every component with text information, picture information, speech information, and structure information which constitute said compound media document accumulated in said memory measure as similarity of the whole compound media document.

[Claim 46]Said similarity setting process as a similarity decision result for said every component, A similarity decision result of text information of said illustrated document, and text information of said document accumulated in said memory measure, A similarity decision result of picture information of an illustrated this document, and picture information of a document accumulated in this memory measure, A similarity decision result of speech information of an illustrated this document, and speech information of a document accumulated in this memory measure, A storage which stored a similar retrieval program of the compound media document according to claim 45 using a similarity decision result of structure information on this illustration **** document, and structure information on a document accumulated in this memory measure.
[Claim 47]Text information, picture information and speech information by which a front similarity setting process is included in said document, And a storage which stored a similar retrieval program of the compound media document according to claim 45 including a linear combination calculation process of making into similarity as said whole compound media document what calculated similarity, applied a value of dignity to this similarity for every characteristic information of structure information, and took linear combination.
[Claim 48]When two or more same media exist in the same document, said document comparison process, About all the search keys for every media containing a text, a picture, and a sound which are contained in said illustrated document. A storage which stored a similar retrieval program of a compound media document given in claims 43 and 44 including a representation similarity determination process that calculate all the similarity to this search key in media in an accumulated document, and this similarity makes the highest thing similarity of representation to said search key.
[Claim 49]A storage which stored a similar retrieval program of the compound media document according to claim 44, comprising:
Said tree with an order label which said similarity calculation process expressed each of structure information on said illustrated document, and structure information on an accumulated document as a tree with an order label (ordered labeledtree) as said search key, and expressed said illustrated document.
A tree shape comparison process with an order label of setting up similarity of structure

information on a document by comparing shape with a tree with an order label expressing said accumulated document.

[Claim 50]A storage which stored a similar retrieval program of the compound media document according to claim 49, comprising:
The number of times which performed editing operation including insertion of a node required in order to change said tree shape comparison process with an order label into a tree showing said document accumulated from a tree which considers that said document structure is a tree and expresses said illustrated document when judging similarity of said structure information, deletion of a node, and change of a node name.
A compilation distance calculation process of setting up compilation distance computed from cost required to perform this editing operation as similarity of said document.

[Claim 51]A storage which stored a similar retrieval program of the compound media document according to claim 44, comprising:
A process of calculating similarity based on characteristic information of text information, picture information and speech information, and structure information that said similarity calculation process is included in a document.
A process of performing the 1st-step selection based on a similarity calculation result based on characteristic information of structure information.
A process of making into similarity as the whole compound media document similarity based on characteristic information of text information and picture information which are contained in said document, or speech information.

[Claim 52]Claim 43 and a storage which stored a similar retrieval program of a compound media document given in 44 characterized by comprising the following.
A process of calculating similarity based on characteristic information of text information, picture information and speech information, and structure information that said similarity calculation process is included in a document.
A process of performing the 1st-step selection based on a similarity calculation result based on characteristic information of said text information, said picture information, or said speech information.
A process of making into similarity as the whole compound media document similarity based on characteristic information of structure information included in said document.

[Claim 53]A storage which stored a similar retrieval program of a compound media document given in claims 43 and 44 which include a document level similarity calculation process of

setting up similarity in a document level of these media when two or more same media exist in a document in which said similarity calculation process was illustrated as a search key.
[Claim 54]A storage which stored a similar retrieval program of the compound media document according to claim 53, comprising:
Said document level similarity calculation process about each search key of two or more said media in a document illustrated as a search key. A process of calculating all the similarity to this search key in these media in an accumulated document about all the search keys for every media including text information, picture information, and speech information which are included in said illustrated document.
A process of making a thing with said highest similarity into similarity of representation to said search key..
A process of calculating average value of similarity of said representation.
A document level similarity setting process which sets up similarity in a document level of said media.

[Claim 55]A storage which stored a similar retrieval program of the compound media document according to claim 54, comprising:
Said document level similarity setting process about each search key of two or more said media in a document illustrated as a search key. A process of calculating all the similarity to this search key in these media in an accumulated document about all the search keys for every media including text information, picture information, and speech information which are included in an illustrated this document.
A process of making a thing with said highest similarity into similarity of representation to said search key.
A process of making into similarity in a document level of said media what has the highest similarity among similarity of said representation.

[Claim 56]Said similarity calculation process expresses each of structure information on a document illustrated as a search key, and structure information on an accumulated document as a tree with an order label (ordered labeled tree), Characteristic information of each media in each document is expressed as a tree with an order label with an attribute stored as an attribute of a node in said tree with an order label, By comparing an attribute and shape with a tree with an order label with an attribute expressing a tree with an order label with an attribute expressing said illustrated document, and said accumulated document. A storage which stored a similar retrieval program of the compound media document according to claim 49 including a similar retrieval process of setting up similarity of a compound statement document.
[Claim 57]Said similar retrieval process about a tree with an order label with an attribute

expressing an accumulated document with a node which has characteristic information which is the attribute of each node of a tree with an order label with an attribute expressing an illustrated document, and similar characteristic information as an attribute. A storage which stored a similar retrieval program of the compound media document according to claim 56 including a process of setting up similarity of a compound media document from a difference in structural physical relationship of this node.

[Claim 58]A storage which stored a similar retrieval program of the compound media document according to claim 49, comprising:

A process of said tree shape comparison process with an order label considering that document structure is a tree with an order label, and evaluating and mapping this characteristic information on multi dimensional vector space based on characteristic information about this tree with an order label.

A process of setting up distance on said vector space as similarity of a document.

[Claim 59]A storage which stored a similar retrieval program of the compound media document according to claim 58 which calculates similarity of a document by evaluating and using position information on a name and a node number of each node of said tree with an order label, or each node as characteristic information about said label tree with an order.

[Claim 60]If it is the frequency of occurrence of a concept which a descriptive content of a text expresses as said characteristic information if it is text information, or each word, and picture information, If it is hue of a picture, chroma saturation, luminosity, color arrangement, and speech information and is strength of a sound, a melody, and structure information, A storage which stored a similar retrieval program of the compound media document according to claim 43 made into characteristic information from which shape of a tree when a tree with an order label expresses document structure, a label name of a node, link information, etc. are extracted from a component of said compound media document.

[Claim 61]A storage which stored a similar retrieval program of the compound media document according to claim 43, comprising:

Information and structure information on media containing a text, a picture, and a sound which constitute said document in which said document comparison process was illustrated considering similarity of an accumulated document to said illustrated document as said search key.

A process of setting up an evaluation value based on a similarity decision result for every component with information on media and structure information containing a text, a picture, and a sound which constitute said document accumulated as whole compound media document similarity.

A process of carrying out ranking by arranging similarity of said accumulated document in a

descending order, and judging similarity.

[Claim 62]A storage which stored a similar retrieval program of the compound media document according to claim 44, comprising:
A process to which said similarity setting process sets similarity for every component of said compound media document.
A process of calculating similarity for every characteristic information of text information included in said document, picture information and speech information, and structure information.
A process of making into similarity as said whole compound media document what applied a value of dignity to said similarity and took linear combination.

[Claim 63]Said similarity setting process as an evaluation value based on a similarity decision result for every component of said document The similarity of each component itself. Or a storage which stored a similar retrieval program of the compound media document according to claim 44 including a process of using what multiplied dignity given to similarity of each component by user.

---

[Translation done.]

* NOTICES *



1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

---

## DETAILED DESCRIPTION

---

[Detailed Description of the Invention]
[0001]
[Field of the Invention]This invention relates to the storage which stored the similar retrieval program of the similar retrieval method of a compound media document, the device, and the compound media document, It is related with the storage which stored the similar retrieval program of the similar retrieval method of the compound media document for searching a similar compound media document about a compound media document data base especially, the device, and the compound media document.
[0002]
[Description of the Prior Art]As a Prior art about a structured document search method, there are combination (Kanamoto, Kato, Kinutani, Yoshikawa, the "structured document searching method in which efficient updating is possible") of StructureIndex of the Nara Institute of Science and Technology study and Content Index, etc. In these systems, indexes, such as an inverted file about the appearing position of the text information stored in structured documents, such as SGML documents and an XML document, or the information on document structure, including element name etc., are prepared beforehand, The combination of text information, or text information and structure information is given as a search key, Search of a structured document is enabled by performing the Boolean retrieval which judges whether the given keyword is contained in the document, and range retrieval which judges whether the keyword given within specified limits appears.
[0003]By systems, such as ConceptBase of JUST System, and VextSearch of Komatsu Soft. The concept included in the text used as the concept included in the text inputted by the natural sentence etc. as a search key and a retrieval object is compared, and search of a similar document (only text) is enabled.
[0004]

[Problem(s) to be Solved by the Invention]However, since the similar retrieval method for a compound media document is not established when it is going to apply the above-mentioned conventional method to the similar retrieval for a compound media document, the following problems occur.

- If a user does not know beforehand the information about document structures, such as an element name (notes: tag name showing structure information) in a structured document, search which gave structure information to a part of search key cannot be performed.

[0005]- Similar retrieval of the compound media document which uses the information and structure information on media other than texts, such as a picture and a sound, as a search key cannot be performed.

If it was made in view of the above-mentioned point and a compound media document is illustrated as a search condition, this invention, Characteristic information included in the illustrated document, such as text information, a picture and audio information, and structure information, and the text information included in the accumulated document, Compare characteristic information, such as a picture, audio information, and structure information, respectively, and The similarity of text information, By calculating the similarity of picture information, the similarity of speech information, and the similarity of structure information separately, multiplying the value of dignity by them, making into the similarity in a document level what calculated the synthetic evaluation value, and adjusting the value of dignity. It aims at providing the storage which stored the similar retrieval program of the similar retrieval method of the compound media document in which the similar retrieval which thought the similarity of text information as important, the similar retrieval which thought the similarity of picture information as important, the similar retrieval which thought the similarity of structure information as important, etc. are possible, the device, and the compound media document.

[0006]

[Means for Solving the Problem]Drawing 1 is a figure for explaining a principle of this invention. This invention (claim 1) is added to text information, picture information, and voice data that is data of human being's voice, In a similar retrieval method of a compound media document which is a structured document which comprises speech information containing composition data containing CD and a record and music data, The 1st characteristic information is extracted from a component of a compound media document which analyzes the syntax of a given compound media document (Step 1), and is obtained as a result of being analyzed (Step 2), Accumulate a document and extracted characteristic information (Step 3), and a compound media document is inputted as a search condition (Step 4), The 2nd characteristic information is extracted from a component of a search condition which analyzes the syntax of an inputted search condition (Step 5), and is acquired as a result of being analyzed (Step 6), Similarity of two compound media documents is judged based on the 1st characteristic information and 2nd

characteristic information that are accumulated (Step 7), and a similar compound media document is outputted (Step 8).

[0007]At the time of search, this invention (claim 2) inputs a document which a user illustrated as a search key, extracts the 2nd characteristic information from an illustrated document, and calculates similarity between documents by the 2nd characteristic information and 1st characteristic information that were extracted. Information and structure information on media containing a text, a picture, and a sound which constitute a document illustrated as a search key when this invention (claim 3) calculates similarity of a compound media document, An evaluation value based on a similarity decision result for every component with information on media and structure information containing a text, a picture, and a sound which constitute a document accumulated is set up as similarity of the whole compound media document.

[0008]Text information of a document in which this invention (claim 4) was illustrated as a similarity decision result for every component, A similarity decision result of text information of an accumulated document, and picture information of an illustrated document, A similarity decision result of a similarity decision result of a similarity decision result with picture information of an accumulated this document, speech information of a this illustrated document, and speech information of a this accumulated document, structure information on a this illustrated document, and structure information on a this accumulated document is used.

[0009]When this invention (claim 5) calculates similarity of a compound media document, for every characteristic information of text information included in a document, picture information and speech information, and structure information, it calculates similarity, applies a value of dignity to similarity, and makes what took linear combination similarity as the whole compound media document.

[0010]When this invention (claim 6) calculates similarity of a compound media document, Text information included in an illustrated document when two or more same media exist in the same document, All the similarity to this search key in these media in an accumulated document is calculated about all the search keys for every media including picture information and speech information, and similarity makes the highest thing similarity of representation to a search key.

[0011]When this invention (claim 7) calculates similarity of a compound media document, as a search key, A tree with an order label which expressed each of structure information on an illustrated document, and structure information on an accumulated document as a tree with an order label (ordered labeled tree), and expressed an illustrated document, Similarity of structure information on a document is set up by comparing shape with a tree with an order label expressing an accumulated document.

[0012]When this invention (claim 8) sets up similarity of structure information, it considers that document structure is a tree, The number of times which performed editing operation including

insertion of a node required in order to change into a tree showing a document accumulated from a tree showing an illustrated document, deletion of a node, and change of a node name, and compilation distance computed from cost required to perform this editing operation are set up as similarity of a document.

[0013]When this invention (claim 9) calculates similarity of a compound media document, Similarity based on characteristic information of text information included in a document, picture information and speech information, and structure information is calculated, The 1st-step selection based on a similarity calculation result based on characteristic information of structure information is performed, and let similarity based on characteristic information of text information and picture information which are contained in a document, or speech information be the similarity as the whole compound media document.

[0014]When this invention (claim 10) calculates similarity of a compound media document, Similarity based on characteristic information of text information included in a document, picture information and speech information, and structure information is calculated, The 1st-step selection based on a similarity calculation result based on characteristic information of text information, picture information, or speech information is performed, and let similarity based on characteristic information of structure information included in a document be the similarity as the whole compound media document.

[0015]This invention (claim 11) sets up similarity in a document level of these media, when calculating similarity of a compound media document and two or more same media exist in a document illustrated as a search key. When this invention (claim 12) sets up similarity in a document level of media, About each search key of two or more media in a document illustrated as a search key. About all the search keys for every media including text information, picture information, and speech information which are included in an illustrated document. All the similarity to this search key in these media in an accumulated document is calculated, and similarity makes the highest thing similarity of representation to a search key, calculates average value of similarity of representation, and sets up similarity in a document level of media.

[0016]When this invention (claim 13) sets up similarity in a document level of media, About each search key of two or more media in a document illustrated as a search key. About all the search keys for every media including text information, picture information, and speech information which are included in an illustrated document. All the similarity to this search key in these media in an accumulated document is calculated, and similarity makes the highest thing similarity of representation to a search key, and makes what has the highest similarity similarity in a document level of media among similarity of representation.

[0017]When this invention (claim 14) performs similar retrieval of a compound media document, Each of structure information on a document illustrated as a search key and

structure information on an accumulated document is expressed as a tree with an order label (ordered labeled tree), Characteristic information of each media in each document is expressed as a tree with an order label with an attribute stored as an attribute of a node in a tree with an order label, Similarity of a compound media document is set up by comparing an attribute and shape with a tree with an order label with an attribute expressing a tree with an order label with an attribute expressing an illustrated document, and an accumulated document.

[0018]When this invention (claim 15) performs similar retrieval of a compound media document, About a tree with an order label with an attribute expressing an accumulated document with a node which has characteristic information which is the attribute of each node of a tree with an order label with an attribute expressing an illustrated document, and similar characteristic information as an attribute. Similarity of a compound media document is set up from a difference in structural physical relationship of a node.

[0019]When it sets up similarity of structure information, this invention (claim 16) considers that document structure is a tree with an order label, based on characteristic information about this tree with an order label, on multi dimensional vector space, evaluates and maps this characteristic information and sets up distance on vector space as similarity of a document. This invention (claim 17) calculates similarity of a document by evaluating and using position information on a name and a node number of each node of a tree with an order label, or each node as characteristic information about a label tree with an order.

[0020]As characteristic information, if this invention (claim 18) is text information, If it is the frequency of occurrence of a concept which a descriptive content of a text expresses, or each word, and picture information, If it is hue of a picture, chroma saturation, luminosity, color arrangement, and speech information and is strength of a sound, a melody, and structure information, let shape of a tree when a tree with an order label expresses document structure, a label name of a node, link information, etc. be the characteristic information from which it is extracted from a component of a compound media document.

[0021]Information and structure information on media containing a text, a picture, and a sound which constitute a document illustrated considering similarity of an accumulated document to an illustrated document as a search key when this invention (claim 19) judged similarity, Ranking is carried out by setting up an evaluation value based on a similarity decision result for every component with information on media and structure information containing a text, a picture, and a sound which constitute a document accumulated as whole compound media document similarity, and arranging similarity of an accumulated document in a descending order, and similarity is judged.

[0022]When this invention (claim 20) sets up similarity, for every component of a compound media document. Similarity is set up, for every characteristic information of text information

included in a document, picture information and speech information, and structure information, similarity is calculated, and a value of dignity is applied to similarity, and let what took linear combination be the similarity as the whole compound media document.

[0023]When this invention (claim 21) sets up similarity of the whole compound media document, it uses the similarity of each component itself, or a thing which multiplied dignity given to similarity of each component by user as an evaluation value based on a similarity decision result for every component of a document. Drawing 2 is a principle lineblock diagram of this invention.

[0024]This invention (claim 22) is provided with the following.

The compound media document input means 10 which is a similarity retrieval device of a compound media document which is a structured document which comprises speech information which adds to text information, picture information, and voice data that is data of human being's voice, and contains composition data containing CD and a record and music data, and inputs a compound media document.

A compound media document given by the compound media document input means 10, and the document analyzing means 40 which analyzes the syntax of an inputted search condition.

A characteristic information extraction means 50 to extract characteristic information from a component of a document obtained as a result of being analyzed by the document analyzing means 40.

The accumulation means 60 which accumulates characteristic information extracted by compound media document and the characteristic information extraction means 50, The search condition input means 30 which inputs a compound media document as a search condition, A document comparison means 80 to judge similarity of two compound media documents based on characteristic information of a compound media document accumulated in the accumulation means 60, and characteristic information extracted by the characteristic information extraction means 50 based on a result of having analyzed an inputted search condition by the document analyzing means 40, The output means 90 which outputs a similar compound media document based on similarity judged by the document comparison means 80.

[0025]In the characteristic information extraction means 50 including a means by which this invention (claim 23) inputs a document which a user illustrated as a search key in the search condition input means 30, An input sentence document characteristic information extraction means to extract a compound media document illustrated as a search key from a given compound media document, In the document comparison means 80 including a search characteristic information extraction means to extract characteristic information of search information from a document illustrated by user, A similarity calculation means to calculate

similarity between compound media documents illustrated as a compound media document and a search key by input sentence document characteristic information extracted by an input sentence document characteristic information extraction means and search characteristic information extracted by a search characteristic information extraction means is included.
[0026]Text information, picture information, speech information, and structure information that this invention (claim 24) constitutes a document illustrated as a search key by the search condition input means 30 in a similarity calculation means, A similarity setting-out means to set up an evaluation value based on a similarity decision result for every component with text information, picture information, speech information, and structure information which constitute a compound media document accumulated in an accumulation means as similarity of the whole compound media document is included.
[0027]In a similarity setting-out means, this invention (claim 25) as a similarity decision result for every component, A similarity decision result of text information of an illustrated document, and text information of a document accumulated in an accumulation means, A similarity decision result of picture information of an illustrated this document, and picture information of a document accumulated in this accumulation means, A similarity decision result of speech information of an illustrated this document and speech information of a document accumulated in this accumulation means and a similarity decision result of structure information on this illustration **** document and structure information on a document accumulated in this accumulation means are used.
[0028]Text information by which this invention (claim 26) is contained in a document in a similarity setting-out means, For every characteristic information of picture information, speech information, and structure information, similarity is calculated, a value of dignity is applied to this similarity, and a linear combination calculating means which makes what took linear combination similarity as the whole compound media document is included. In a document comparison means, when two or more same media exist in the same document, this invention (claim 27), All the similarity to this search key in media in an accumulated document is calculated about all the search keys for every media containing a text, a picture, and a sound which are contained in an illustrated document, and this similarity contains a representation similarity determination means which makes the highest thing similarity of representation to a search key.
[0029]In a similarity calculation means, this invention (claim 28) as a search key, A tree with an order label which expressed each of structure information on an illustrated document, and structure information on an accumulated document as a tree with an order label (ordered labeled tree), and expressed an illustrated document, By comparing shape with a tree with an order label expressing an accumulated document, a tree shape comparison means with an order label to set up similarity of structure information on a document is included.

[0030]In a tree shape comparison means with an order label, when this invention (claim 29) judges similarity of structure information, The number of times which performed editing operation including insertion of a node required in order to change into a tree showing a document accumulated from a tree which considers that document structure is a tree and expresses an illustrated document, deletion of a node, and change of a node name, A compilation distance calculating means which sets up compilation distance computed from cost required to perform this editing operation as similarity of a document is included.
[0031]This invention (claim 30) is provided with the following.
A means to calculate similarity based on characteristic information of text information included in a document, picture information and speech information, and structure information in a similarity calculation means.
A means to perform the 1st-step selection based on a similarity calculation result based on characteristic information of structure information.
A means which makes similarity based on characteristic information of text information and picture information which are contained in a document, or speech information similarity as the whole compound media document.

[0032]This invention (claim 31) is provided with the following.
A means to calculate similarity based on characteristic information of text information included in a document, picture information and speech information, and structure information in a similarity calculation means.
A means to perform the 1st-step selection based on a similarity calculation result based on characteristic information of text information, picture information, or speech information.
A means which makes similarity based on characteristic information of structure information included in a document similarity as the whole compound media document.

[0033]In a similarity calculation means, this invention (claim 32) contains a document level similarity calculation means to set up similarity in a document level of these media, when two or more same media exist in a document illustrated as a search key. This invention (claim 33) is provided with the following.
About each search key of two or more media in a document illustrated as a search key in a document level similarity calculation means. A means to calculate all the similarity to this search key in these media in an accumulated document about all the search keys for every media including text information, picture information, and speech information which are included in an illustrated document.
A means by which similarity makes the highest thing similarity of representation to a search key.

A means to calculate average value of similarity of representation.
A document level similarity setting-out means to set up similarity in a document level of media.

[0034]This invention (claim 34) is provided with the following.
About each search key of two or more media in a document illustrated as a search key in a document level similarity setting-out means. A means to calculate all the similarity to this search key in these media in an accumulated document about all the search keys for every media including text information, picture information, and speech information which are included in an illustrated this document.
A means by which similarity makes the highest thing similarity of representation to a search key.
A means which makes what has the highest similarity similarity in a document level of media among similarity of representation.

[0035]This invention (claim 35) expresses each of structure information on a document illustrated as a search key, and structure information on an accumulated document as a tree with an order label (ordered labeled tree) in a similarity calculation means, Characteristic information of each media in each document is expressed as a tree with an order label with an attribute stored as an attribute of a node in a tree with an order label, By comparing an attribute and shape with a tree with an order label with an attribute expressing a tree with an order label with an attribute expressing an illustrated document, and an accumulated document, a similar retrieval means to set up similarity of a compound media document is included.
[0036]This invention (claim 36) about a tree with an order label with an attribute expressing an accumulated document with a node which has characteristic information which is the attribute of each node of a tree with an order label with an attribute which expressed an illustrated document in a similar retrieval means, and similar characteristic information as an attribute. A means to set up similarity of a compound media document from a difference in structural physical relationship of this node is included.
[0037]This invention (claim 37) is provided with the following.
A means to consider that document structure is a tree with an order label, and to evaluate and map this characteristic information on multi dimensional vector space in a tree shape comparison means with an order label based on characteristic information about this tree with an order label.
A means to set up distance on vector space as similarity of a document.

[0038]This invention (claim 38) calculates similarity of a document by evaluating and using

position information on a name and a node number of each node of a tree with an order label, or each node as characteristic information about a label tree with an order. As characteristic information, if this invention (claim 39) is text information, If it is the frequency of occurrence of a concept which a descriptive content of a text expresses, or each word, and picture information, If it is hue of a picture, chroma saturation, luminosity, color arrangement, and speech information and is strength of a sound, a melody, and structure information, let shape of a tree when a tree with an order label expresses document structure, a label name of a node, link information, etc. be the characteristic information from which it is extracted from a component of a compound media document.

[0039]This invention (claim 40) is provided with the following.

Information and structure information on media containing a text, a picture, and a sound which constitute a document illustrated considering similarity of an accumulated document to an illustrated document as a search key in the document comparison means 80.

A means to set up an evaluation value based on a similarity decision result for every component with information on media and structure information containing a text, a picture, and a sound which constitute a document accumulated as whole compound media document similarity.

A means to carry out ranking by arranging similarity of an accumulated document in a descending order, and to judge similarity.

[0040]This invention (claim 41) is provided with the following.

A means to set up similarity for every component of a compound media document in a similarity setting-out means.

A means to calculate similarity for every characteristic information of text information included in a document, picture information and speech information, and structure information.

A means which makes what applied a value of dignity to similarity and took linear combination similarity as the whole compound media document.

[0041]This invention (claim 42) contains a means to use the similarity of each component itself, or a thing which multiplied dignity given to similarity of each component by user as an evaluation value based on a similarity decision result for every component of a document, in a similarity setting-out means. This invention (claim 43) is provided with the following.

It adds to text information, picture information, and voice data that is data of human being's voice, A compound media document which is the storage which stored a similar retrieval program of a compound media document which is a structured document which comprises speech information containing composition data containing CD and a record and music data, and was given, and a document analyzing process of analyzing the syntax of an inputted

search condition.
A characteristic information extraction process of extracting characteristic information from a component of a document obtained as a result of being analyzed in a document analyzing process.
A storing process of storing in a memory measure characteristic information extracted in a compound media document and a characteristic information extraction process.
A search condition input process of making a compound media document inputting as a search condition, A document comparison process of judging similarity of two compound media documents from characteristic information of a compound media document accumulated in a memory measure, and characteristic information extracted in a characteristic information extraction process based on a result of having analyzed an inputted search condition in a document analyzing process, An output process to which a similar compound media document is made to output based on similarity judged in a document comparison process.

[0042]In a characteristic information extraction process including a process as which this invention (claim 44) inputs a document which a user illustrated as a search key in a search condition input process, An input sentence document characteristic information extraction process of extracting characteristic information from a given compound media document, In a document comparison process including a search characteristic information extraction process of extracting characteristic information of a compound media document illustrated as a search key from a document illustrated by user, A similarity calculation process of calculating similarity between compound media documents illustrated as a compound media document and a search key by input sentence document characteristic information extracted in an input sentence document characteristic information extraction process and search characteristic information extracted in a search characteristic information extraction process is included.
[0043]Text information, picture information, speech information, and structure information that this invention (claim 45) constitutes a document illustrated as a search key in a search condition input process in a similarity calculation process, A similarity setting process which sets up an evaluation value based on a similarity decision result for every component with text information, picture information, speech information, and structure information which constitute a compound media document accumulated in a memory measure as similarity of the whole compound media document is included.
[0044]In a similarity setting process, this invention (claim 46) as a similarity decision result for every component, A similarity decision result of text information of an illustrated document, and text information of a document accumulated in a memory measure, A similarity decision result of picture information of an illustrated this document, and picture information of a document

accumulated in this memory measure, A similarity decision result of speech information of an illustrated this document and speech information of a document accumulated in this memory measure and a similarity decision result of structure information on this illustration **** document and structure information on a document accumulated in this memory measure are used.

[0045]Text information by which this invention (claim 47) is contained in a document in a similarity setting process, For every characteristic information of picture information, speech information, and structure information, similarity is calculated, a value of dignity is applied to this similarity, and a linear combination calculation process of making into similarity as the whole compound media document what took linear combination is included. In a document comparison process, when two or more same media exist in the same document, this invention (claim 48), About all the search keys for every media containing a text, a picture, and a sound which are contained in an illustrated document. All the similarity to this search key in media in an accumulated document is calculated, and this similarity includes a representation similarity determination process of making the highest thing into similarity of representation to a search key.

[0046]In a similarity calculation process, this invention (claim 49) as a search key, A tree with an order label which expressed each of structure information on an illustrated document, and structure information on an accumulated document as a tree with an order label (ordered labeled tree), and expressed an illustrated document, By comparing shape with a tree with an order label expressing an accumulated document, a tree shape comparison process with an order label of setting up similarity of structure information on a document is included.

[0047]In a tree shape comparison process with an order label this invention (claim 50), The number of times which performed editing operation including insertion of a node required [ when judging similarity of structure information ] in order to change into a tree showing a document accumulated from a tree which considers that document structure is a tree and expresses an illustrated document, deletion of a node, and change of a node name, A compilation distance calculation process of setting up compilation distance computed from cost required to perform this editing operation as similarity of a document is included.

[0048]This invention (claim 51) is provided with the following.

A process of calculating similarity based on characteristic information of text information included in a document, picture information and speech information, and structure information in a similarity calculation process.

A process of performing the 1st-step selection based on a similarity calculation result based on characteristic information of structure information.

A process of making into similarity as the whole compound media document similarity based on characteristic information of text information and picture information which are contained in

a document, or speech information.

[0049]This invention (claim 52) is provided with the following.
A process of calculating similarity based on characteristic information of text information included in a document, picture information and speech information, and structure information in a similarity calculation process.
A process of performing the 1st-step selection based on a similarity calculation result based on characteristic information of text information, picture information, or speech information.
A process of making into similarity as the whole compound media document similarity based on characteristic information of structure information included in a document.

[0050]In a similarity calculation process, this invention (claim 53) includes a document level similarity calculation process of setting up similarity in a document level of these media, when two or more same media exist in a document illustrated as a search key. This invention (claim 54) is provided with the following.
About each search key of two or more media in a document illustrated as a search key in a document level similarity calculation process. A process of calculating all the similarity to this search key in these media in an accumulated document about all the search keys for every media including text information, picture information, and speech information which are included in an illustrated document.
A process of making into similarity of representation to a search key what has the highest similarity.
A process of calculating average value of similarity of representation.
A document level similarity setting process which sets up similarity in a document level of media.

[0051]This invention (claim 55) is provided with the following.
About each search key of two or more media in a document illustrated as a search key in a document level similarity setting process. A process of calculating all the similarity to this search key in these media in an accumulated document about all the search keys for every media including text information, picture information, and speech information which are included in an illustrated this document.
A process of making into similarity of representation to a search key what has the highest similarity.
A process of making into similarity in a level in a media document what has the highest similarity among similarity of representation.

[0052]This invention (claim 56) expresses each of structure information on a document illustrated as a search key, and structure information on an accumulated document as a tree with an order label (ordered labeled tree) in a similarity calculation process, Characteristic information of each media in each document is expressed as a tree with an order label with an attribute stored as an attribute of a node in a tree with an order label, By comparing an attribute and shape with a tree with an order label with an attribute expressing a tree with an order label with an attribute expressing an illustrated document, and an accumulated document, a similar retrieval process of setting up similarity of a compound media document is included.

[0053]This invention (claim 57) about a tree with an order label with an attribute expressing an accumulated document with a node which has characteristic information which is the attribute of each node of a tree with an order label with an attribute which expressed an illustrated document in a similar retrieval process, and similar characteristic information as an attribute. A process of setting up similarity of a compound media document from a difference in structural physical relationship of this node is included.

[0054]This invention (claim 58) is provided with the following.

A process of considering that document structure is a tree with an order label, and evaluating and mapping this characteristic information on multi dimensional vector space in a tree shape comparison process with an order label based on characteristic information about this tree with an order label.

A process of setting up distance on vector space as similarity of a document.

[0055]This invention (claim 59) calculates similarity of a document by evaluating and using position information on a name and a node number of each node of a tree with an order label, or each node as characteristic information about a label tree with an order. As characteristic information, if this invention (claim 60) is text information, If it is the frequency of occurrence of a concept which a descriptive content of a text expresses, or each word, and picture information, If it is hue of a picture, chroma saturation, luminosity, color arrangement, and speech information and is strength of a sound, a melody, and structure information, let shape of a tree when a tree with an order label expresses document structure, a label name of a node, link information, etc. be the characteristic information from which it is extracted from a component of a compound media document.

[0056]This invention (claim 61) is provided with the following.

Information and structure information on media containing a text, a picture, and a sound which constitute a document illustrated considering similarity of an accumulated document to an illustrated document as a search key in a document comparison process.

A process of setting up an evaluation value based on a similarity decision result for every

component with information on media and structure information containing a text, a picture, and a sound which constitute a document accumulated as whole compound media document similarity.
A process of carrying out ranking by arranging similarity of an accumulated document in a descending order, and judging similarity.

[0057]This invention (claim 62) is provided with the following.
A process of setting up similarity for every component of a compound media document in a similarity setting process.
A process of calculating similarity for every characteristic information of text information included in a document, picture information and speech information, and structure information.
A process of making into similarity as the whole compound media document what applied a value of dignity to similarity and took linear combination.

[0058]This invention (claim 63) includes a process of using the similarity of each component itself, or a thing which multiplied dignity given to similarity of each component by user as an evaluation value based on a similarity decision result for every component of a document, in a similarity setting process. As mentioned above, in this invention, it becomes possible to determine a unit of components, such as information on media, and structure information, which should compare between an illustrated document and an accumulated document by analyzing the syntax of a given document.
[0059]Search based on the contents of the document or information on the logical structure is enabled by a document being characterized by extracted characteristic information. It becomes possible [ not only text information but picture information, speech information, structure information, etc. ] to use as a part of search key. It becomes possible to access a document at high speed by creating an index from text information accumulated, picture information, speech information, structure information, etc. to a document having contained them.
[0060]Since the contents of characteristic information, such as text information of a document and picture information, can be checked on a display, it is easy to input a compound media document including characteristic information which a user meant as a search key. For every characteristic information included in a compound media document inputted as a search key, similarity is calculated and an evaluation value based on them is calculated. For example, for every characteristic information of a document, similarity is calculated and similar retrieval of a compound media document which also used the similarity of a picture, speech information, and structure information as a search condition in addition to text information becomes possible by calculating what applied, added and united a value of dignity with them as similarity in a document level.

[0061]Since similarity is calculated for every characteristic information in a document, a similarity calculation method that only similarity calculation methods of picture information differ can be adopted concerning each similarity calculation method, and replacing selectively can carry out easily.

[0062]

[Embodiment of the Invention]As a component of a compound media document, as shown in drawing 3, there are text information, picture information, speech information, structure information, etc. Hereafter, the similar retrieval in the compound media document concerned is explained. Drawing 4 shows the composition of the similarity retrieval device of the compound media document of this invention.

[0063]The similarity retrieval device of the compound media document shown in the figure comprises the compound media document input device 10, the search condition input device 20, the search condition input section 30, the compound media document analyzing part 40, the characteristic information extraction part 50, the accumulating part 60, the memory 70, the document comparing element 80, and the display 90. The compound media document input device 10 inputs a document including text information, picture information, speech information, and structure information.

[0064]The search condition input devices 20 are pointing devices, such as a mouse used for a user's input, a keyboard, etc. The search condition input section 30 makes a user input a document filing name from the keyboard which is the search condition input device 20, or. The compound media document inputted as a search key by making a mouse operate it, and making it click on the icon of a document, or making the document obtained by the last search results click with a mouse is acquired. The compound media document which serves as a search key for searching a compound media document in detail is illustrated. A search key is illustrated by specifying the document filing name to illustrate or clicking on the icon of the document to illustrate on a display with a pointing device etc. It is possible to specify the portion than to which a user wants to attach greater importance to similarity, when illustrating a document, and it is possible to change the value of the dignity which shows the degree to think as important suitably, and to input it to the characteristic information of the portion which wants to think similarity as important. At this time, the document number k returned as the value and search results of the dignity of which portion in a document to think the similarity as important is acquired from a user. Or the default value of a system is used.

[0065]The compound media document analyzing part 40 analyzes the syntax of the compound media document input device 10 or the document given from the search condition input section 30, and detects the component of documents, such as text information, picture information, speech information, and structure information. The compound media document analyzing part 40 analyzes the document inputted using the purser (parser: syntax analyzer) of SGML or

XML, and detects the component of documents, such as text information, picture information, speech information, and structure information, from a document here.

[0066]The characteristic information extraction part 50 extracts the characteristic information expressing the feature of the component of documents, such as text information, picture information, speech information, and structure information. For example, if the concept etc. which the descriptive content of a text expresses if it is text information are picture information, and the hue of picture information, chroma saturation, luminosity, color arrangement, etc. are speech information, characteristic information, such as strength of a sound and a melody, is extracted with ID of a document, and the element name and the information on an appearing position that characteristic information was stored. If it is structure information, it is considered as the characteristic information from which the shape (layered structure etc.) of a tree when a tree with an order label expresses document structure, the label name of a node, link information, etc. are extracted from the component of a compound media document.

[0067]The accumulating part 60 accumulates the given document in the memory 70. The index to the document which included the characteristic information concerned from each characteristic information is created. By comparing the characteristic information of the illustrated compound media document and the compound media document accumulated in the memory 70, the document comparing element 80 asks for similarity, and outputs the high thing of similarity. The similarity as a compound media document should calculate the evaluation value based on each similarity calculation results, such as text information, picture information, speech information, and structure information. Concerning [ for example, ] similarity, such as text information, picture information, and speech information, Based on a multi dimensional vector space model, map each characteristic information to up to multi dimensional vector space, and if the distance for two points of the characteristic information of the illustrated document on multi dimensional vector space and the characteristic information of the accumulated document is near, It is possible to adopt the approach of setting up so that similarity may become high. It is also possible to carry out ranking by arranging the similarity of the accumulated document in a descending order, and to judge similarity.

[0068]Hereafter, the operation in the above-mentioned composition is divided into a compound media document storing phase and a compound media document-retrieval phase, and is explained. Drawing 5 is a flow chart of the compound media document storing phase of this invention.

Step 101 A compound media document is first inputted from the compound media document input device 10.

[0069]Step 102 The compound media document analyzing part 40 analyzes the syntax of the compound media document inputted from the compound media document input device 10, and detects the component of documents, such as text information, picture information, speech

information, and structure information.
Step 103, next the characteristic information extraction part 50, About document constituents, such as text information, picture information, speech information, and structure information, if it is text information, for example, If the concept etc. which the descriptive content of a text expresses are picture information and the hue, chroma saturation and luminosity of a picture, color arrangement, etc. are speech information, characteristic information, such as strength of a sound and a melody, is extracted with ID of a document, and the element name and the information on an appearing position that characteristic information was stored. The processing concerned is repeated several minutes of all the components.
[0070]Step 104 The accumulating part 60 creates the index to the given document and the document which included the characteristic information concerned from each characteristic information, and stores it in the memory 70. Next, operation of a compound media document-retrieval phase is explained. Drawing 6 is a flow chart of the compound media document-retrieval phase of this invention.
Step 201 the search condition input section 30 makes a document filing name input from the keyboard which is the search condition input device 20, or. The compound media document inputted as a search key is acquired by making a mouse operate it, and making it click on the icon of a document, or making the document obtained by the last search results click with a mouse. At this time, the value of the dignity of which portion in a document to think the similarity as important, and the document number k returned as search results are acquired from a user. Or the default value of a system is used.
[0071]Step 202, next the compound media document analyzing part 40, The syntax of the compound media document inputted from the compound media search condition input section 30 is analyzed like processing of a compound media document storing phase, and the component of documents, such as text information, picture information, speech information, and structure information, is detected.
Step 203 the characteristic information extraction part 50 like a compound media document storing phase, The characteristic information of document constituents, such as text information, picture information, speech information, and structure information, It extracts with ID of a document, and the element name and the information on an appearing position that characteristic information was stored, and characteristic information is extracted about the component of documents, such as text information of the illustrated document, picture information, speech information, and structure information. The processing concerned is repeated several minutes of all the components.
[0072]Step 204 The document comparing element 80 compares the characteristic information of the illustrated document with the characteristic information of the document accumulated in the memory 70, calculates similarity for each characteristic information of every, and calculates

the evaluation value based on those calculation results as similarity as a compound media document. The calculation method of similarity is mentioned later.
Step 205 The document comparing element 80 arranges similarity in a descending order, and chooses from an index the document of top k affairs which the user demanded as a high document of similarity.
[0073]Step 206 It displays on the display 90 by making the high document of the selected similarity into search results. Next, how to ask for the similarity in the above is explained. Drawing 7 is a flow chart at the time of performing document comparison for asking for the similarity of this invention (the 1).
[0074]Step 301 The document comparing element 80 acquires the characteristic information of the document (search condition) inputted from the search condition input section 30, and the characteristic information of the compound media document inputted from the accumulating part 60 from the characteristic information extraction part 50.
Step 302 When characteristic information is structure information, it shifts to Step 303, and when that is not right, it shifts to Step 304.
[0075]Step 303 The document comparing element 80 calculates the compilation distance between the two trees concerned by considering that the structure information on the characteristic information of a search condition and the characteristic information of a compound media document is a tree, and shifts to Step 306. The methods of calculating the structural physical relationship between nodes, such as multi-dimensional-vector-izing a tree and carrying out [ which it is on multi dimensional vector space ] distance calculation, are also possible.
[0076]Step 304 The document comparing element 80 calculates the distance on multi dimensional vector space.
Step 305 Distance chooses the minimum thing as a representative among characteristic information of the same kind.
Step 306 The similarity in a document level is calculated.
[0077]Step 307 Similarity chooses a high document from an index.
Step 308 The selected document is outputted to the display 90. The thing of the above-mentioned similarity calculation searched for for similarity as follows as law on the other hand, for example is possible.
The 1st similarity calculation method : (1) Information and structure information on media containing the text, picture, and sound which constitute the document illustrated as a search key, The evaluation value based on the similarity decision result for every component with the information on media and structure information containing the text, picture, and sound which constitute the document accumulated in the memory 70 is set up as whole compound media document similarity. Here, with the evaluation value based on a similar decision result, the

similarity of each component itself or the thing which multiplied the dignity given to the similarity of each component by the user is used.

[0078](2) The 2nd similarity calculation method : calculate the similarity of text information by finding the distance on the multi dimensional vector space of the characteristic information of the inputted text information, and the characteristic information of the text information accumulated in the memory 70. The similarity of picture information is calculated by finding the distance on the multi dimensional vector space of the characteristic information of the inputted picture information, and the characteristic information of the picture information accumulated in the memory 70.

[0079]The similarity of speech information is calculated by finding the distance on the multi dimensional vector space of the characteristic information of the inputted speech information, and the characteristic information of the speech information accumulated in the memory 70. In each above-mentioned information, what has a small distance on multi dimensional vector space is calculated as what has high similarity.

[0080](3) similarity calculation method [ of ** a 3rd ]: -- the above-mentioned method of (2) again -- in addition, when two or more same media exist in the same document, similarity sets to the similarity of representation [ the highest thing ]. For example, as shown in drawing 8, in the document containing two or more picture information, into the accumulated document, similarity with two or more picture information is calculated, and similarity sets up the highest thing as similarity of representation in it about the picture information which exists in the document illustrated as a search key. This is performed about all the picture information which exists in the document illustrated as a search key.

[0081]In drawing 8, it asks for each similarity of the picture A in an illustration document, and the picture a, b, and c in a stored document, and similarity makes the highest thing (for example, the picture a) the similar picture in the stored document to the picture A in an illustration document. It asks for each similarity of the picture B in an illustration document, and the picture a, b, and c in a stored document, and similarity makes the highest thing (for example, the picture c) the similar picture in the stored document to the picture B in an illustration document.

[0082](4) The 4th similarity calculation method : the structure information of the characteristic information considers that document structure is a tree, calculates compilation distance required in order to change into another tree from one tree, and if compilation distance is small, it will set it up again so that similarity may become high. Compilation distance is computed from cost required to perform [ the number of times which performed editing operation called insertion of a required node, deletion of a node, and change of a node name when changing a tree, and ] those editing operation. Thereby, what has compilation distance possible calculating similarity and small is calculated as a high thing of similarity.

[0083]The 5th similarity calculation method : (5) Text information, picture information, speech information, The value of the dignity of which portion in the document which calculated similarity, such as structure information, respectively and was acquired by the search condition input section 30 to think the similarity as important, Or based on the default value of a system, the value of the individual dignity given to each of similarity, such as text information, picture information, speech information, and structure information, is applied, and linear combination is taken. What took this linear combination is equivalent to the similarity as a compound media document.

[0084](6) In 6th similarity calculation method:, next the document comparing element 80, when calculating the similarity of a compound media document, After calculating the similarity based on the characteristic information of the text information included in a document, picture information and speech information, and structure information and performing the 1st-step selection based on the similarity calculation result based on the characteristic information of structure information, Let similarity based on the characteristic information of the text information and picture information which are contained in a document, or speech information be the similarity as the whole compound media document.

[0085]This method is explained in detail below. Drawing 9 is a flow chart at the time of performing document comparison for asking for the similarity of this invention (the 2).

Step 401 In the characteristic information information extraction part 50, the characteristic information over the inputted search condition is inputted.

[0086]The similarity of step 402 text information, It calculates by finding the distance on the multi dimensional vector space of the characteristic information of the inputted text information, and the characteristic information of the text information accumulated in the memory 70, or searching for the difference of the value obtained from the frequency of occurrence etc. of the characteristic information of the inputted text information, and the characteristic information of the accumulated text information.

[0087]The similarity of picture information is calculated by finding the distance on the multi dimensional vector space of the characteristic information of the inputted picture information, and the characteristic information of the picture information accumulated in the memory 70. The similarity of speech information is calculated by finding the distance on the multi dimensional vector space of the characteristic information of the inputted speech information, and the characteristic information of the speech information accumulated in the memory 70. What has a small distance on multi dimensional vector space, and what has a small absolute value of the difference of the value obtained from the frequency of occurrence etc. are calculated as what has high similarity.

[0088]What [ compilation distance required in order that structure information may consider that document information is a tree and may change it into another tree from one tree among

characteristic information is calculated for ], It is possible to calculate similarity by evaluating and multi-dimensional-vector-izing wooden characteristic information, and finding the distance on multi dimensional vector space etc. What has small compilation distance, and what has a small distance on multi dimensional vector space are calculated as a high thing of similarity.
[0089]Similarity, such as step 403 text information, picture information, speech information, and structure information, is calculated, respectively, Based on the value of the dignity of which portion in the document acquired by the search condition input section 30 to think the similarity as important, or the default value of a system, the 1st-step selection based on the similarity of structure information is performed.
Step 404 Similarity, such as text information of the document which remained after performing the 1st-step selection, picture information, and speech information, is equivalent to the similarity as a compound media document.
[0090]Step 405 Similarity chooses a high document from an index.
Step 406 The selected document is displayed on the display 90.
(7) In similarity calculation method [ of ** a 7th ];, next the document comparing element 80, when calculating the similarity of a compound media document, After calculating the similarity based on the characteristic information of the text information included in a document, picture information and speech information, and structure information and performing the 1st-step selection based on the similarity calculation result based on the characteristic information of text information, picture information, or speech information, Let similarity based on the characteristic information of the structure information included in a document be the similarity as the whole compound media document.
[0091]This method is explained in detail below. Drawing 10 is a flow chart at the time of performing document comparison for asking for the similarity of this invention (the 3).
Step 501 The characteristic information of the inputted search condition is inputted.
The similarity of step 502 text information, It calculates by finding the distance on the multi dimensional vector space of the characteristic information of the inputted text information, and the characteristic information of the text information accumulated in the memory 70, or searching for the difference of the value obtained from the frequency of occurrence etc. of the characteristic information of the inputted text information, and the characteristic information of the accumulated text information.
[0092]The similarity of picture information is calculated by finding the distance on the multi dimensional vector space of the characteristic information of the inputted picture information, and the characteristic information of the accumulated picture information. The similarity of speech information is calculated by finding the distance on the multi dimensional vector space of the characteristic information of the inputted speech information, and the characteristic information of the accumulated speech information. What has a small distance on multi

dimensional vector space, and what has a small absolute value of the difference of the value obtained from the frequency of occurrence etc. are calculated as what has high similarity.
[0093]What [ compilation distance required in order that structure information may consider that document structure is a tree and may change it into another tree from one tree among characteristic information is calculated for ], It is possible to calculate similarity by evaluating and multi-dimensional-vector-izing wooden characteristic information, and finding the distance on multi dimensional vector space etc. What has small compilation distance, and what has a small distance on multi dimensional vector space are calculated as what has high similarity.
[0094]Similarity, such as step 503 text information, picture information, speech information, and structure information, is calculated, respectively, Based on the value of the dignity of which portion in the document acquired by the search condition input section 30 to think the similarity as important, or the default value of a system, the 1st-step selection based on similarity, such as text information, picture information, and speech information, is performed.
[0095]Step 504 After performing the 1st-step selection, the similarity of the structure information on the remaining document is equivalent to the similarity as a compound media document.
Step 505 The high document of similarity is chosen from an index.
Step 506 The selected document is outputted to the display 90.
(8) The 8th similarity calculation method : when calculating the similarity of a compound media document and two or more same media exist in the document illustrated as a search key, set up the similarity in the document level of the media concerned. For example, by (8), although above (3) described the case where two or more same media existed in the document of the accumulated retrieval object, when three different pictures in the document used as a search key exist, the point how to set up the similarity of the picture in the document used as a search key is explained.
[0096]When calculating the similarity of a compound media document and two or more same media exist in the document illustrated as a search key, based on two kinds of examples which set up the similarity in the document level of the media concerned, the case where picture information exists in a document is described.
** When setting up the similarity in the document level of media, about each search key of two or more media concerned in the document illustrated as a search key. All the similarity to the search key in the media in the accumulated document is calculated, When similarity makes the highest thing the similarity of representation to a search key, calculates the average value of the similarity of representation and considers it as the similarity in the document level of media, the picture a is acquired as a similar picture in the stored document to the picture A in an illustration document shown in drawing 8. The picture c is acquired as a similar picture in the stored document to the picture B in an illustration document. The average value of the

similarity between the picture A and the picture a and the similarity between the picture B and the picture c is calculated, and the value is set to the similarity of the picture information in the document level between an illustration document and a stored document.

[0097]** When setting up the similarity in the document level of media, about each search key of two or more media concerned in the document illustrated as a search key. All the similarity to the search key in the media in the accumulated document is calculated, When similarity makes the highest thing the similarity of representation to a search key and makes what has the highest similarity the similarity in the document level of media among the similarity of representation, the picture a is acquired as a similar picture in the stored document to the picture A in an illustration document shown in drawing 8. The picture c is acquired as a similar picture in the stored document to the picture B in an illustration document. What has the highest similarity (for example, similarity between the picture A and the picture a) is set up as similarity of the picture information in the document level between an illustration document and a stored document among the similarity between the picture A and the picture a, and the similarity between the picture B and the picture c.

[0098]Next, when performing similar retrieval of a compound media document, the example which expresses each of the structure information on the document illustrated as a search key and the structure information on the accumulated document as a tree with an order label (ordered labeled tree) is explained. Drawing 11 is a figure for explaining expressing the compound media document of this invention as a tree with an order label with an attribute.

[0099]When performing similar retrieval of a compound media document, it expresses as a tree with an order label with an attribute (tree which extended the tree with an order label) which stored the characteristic information of each media in each document as an attribute of the node in the tree with an order label concerned, The similarity of a compound media document is set up by comparing an attribute and shape with the tree with an order label with an attribute expressing the tree with an order label with an attribute expressing the illustrated document, and the accumulated document.

[0100]When setting up the similarity of this compound media document, based on the characteristic information about a tree with a label with an order, on multi dimensional vector space, the characteristic information concerned is evaluated and mapped and the distance on the vector space concerned is set up as similarity of a document. Evaluation of characteristic information shall evaluate the name (label name) of each node, and a node number and the position information on each node.

[0101]As mentioned above, even if a user does not know detailed document structure, search of the document which also used structure information in addition to text information can carry out. The information and structure information on a picture or a sound can also be included in a part of search key besides text information, and similar retrieval of a document can be

performed. The search condition input section 30, the compound media document analyzing part 40, the characteristic information extraction part 50, the accumulating part 60, and the document comparing element 80 which are shown in drawing 3 are built as a program, It stores in portable storages connected to the computer used as a similarity retrieval device of a compound media document, such as a disk unit, a floppy disk, and CD-ROM, and this invention can be easily realized by installing, when carrying out this invention.

[0102]Change and application are variously possible for this invention within a claim, without being limited to the above-mentioned example.

[0103]

[Effect of the Invention]As mentioned above, according to this invention, even if a user does not know detailed document structure, the document also using structure information other than text information can be searched. The information and structure information on a picture or a sound can also be included in a part of search key besides text information, and resemblance of a document can be searched.

---

[Translation done.]

* NOTICES *



1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

---

## DESCRIPTION OF DRAWINGS

---

[Brief Description of the Drawings]
[Drawing 1]It is a figure for explaining the principle of this invention.
[Drawing 2]It is a principle lineblock diagram of this invention.
[Drawing 3]It is a figure for explaining the compound media document of this invention.
[Drawing 4]It is a lineblock diagram of the similarity retrieval device of the compound media document of this invention.
[Drawing 5]It is a flow chart of the compound media document storing phase of this invention.
[Drawing 6]It is a flow chart of the compound media document-retrieval phase of this invention.

[Drawing 7]It is a flow chart at the time of performing document comparison for asking for the similarity of this invention (the 1).
[Drawing 8]It is a figure for explaining the method of similarity setting out in case two or more same media exist in the same document of this invention.
[Drawing 9]It is a flow chart at the time of performing document comparison for asking for the similarity of this invention (the 2).
[Drawing 10]It is a flow chart at the time of performing document comparison for asking for the similarity of this invention (the 3).
[Drawing 11]It is a figure for explaining expressing as a tree with an order label with an attribute of the compound media document of this invention.
[Description of Notations]
10 A compound media document input means, a compound media document input device
20 Search condition input device
30 A search condition input means, a search condition input section
40 A document analyzing means, a compound media document analyzing part
50 A feature extraction means, a characteristic information extraction part

60 An accumulation means, an accumulating part

70 Memory

80 A document comparison means, a document comparing element

90 An output means, a display

---

[Translation done.]

* NOTICES *



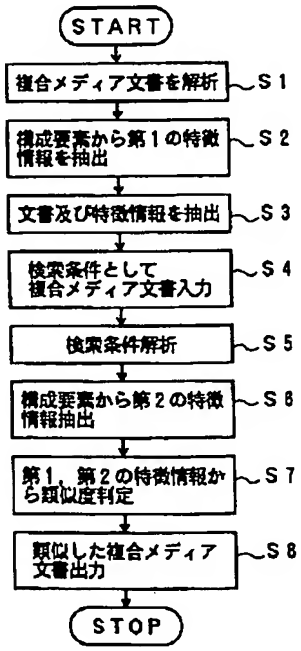
1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

---

## DRAWINGS

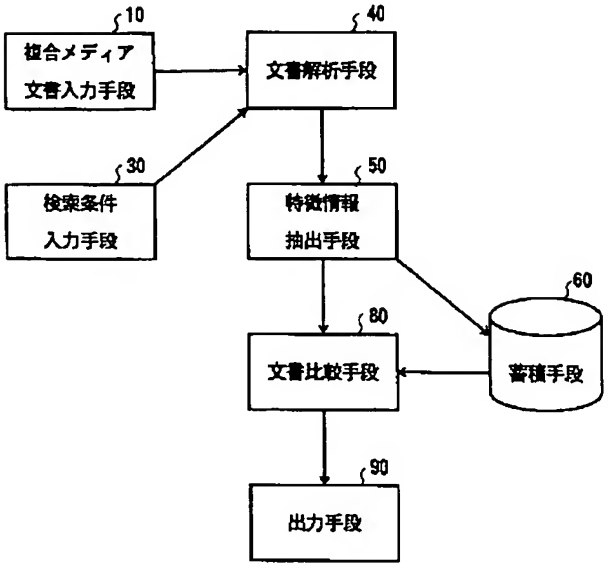
---

[Drawing 1]
本発明の原理を説明するための図



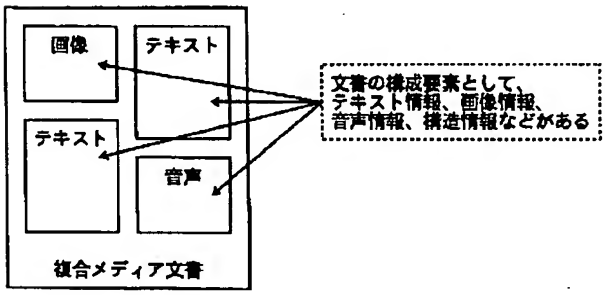

[Drawing 2]

本発明の原理構成図



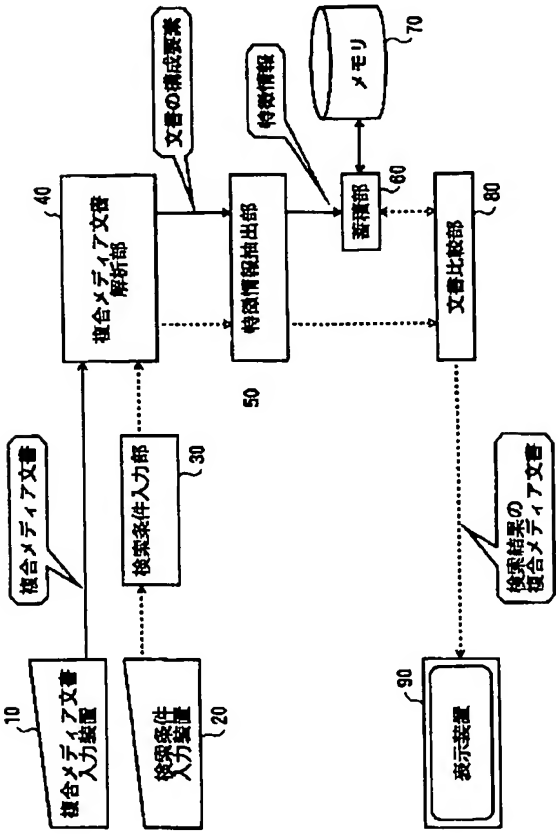

[Drawing 3]

本発明の複合メディア文書を説明するための図



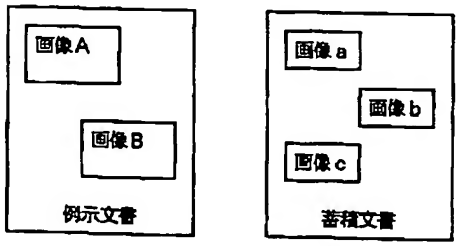

[Drawing 4]

本発明の複合メディア文書の類似検索装置の構成図



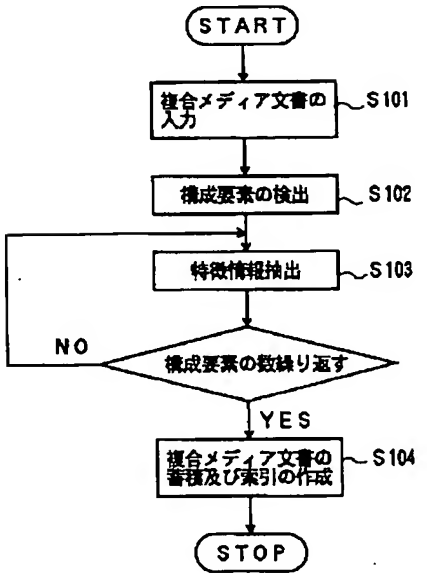

[Drawing 8]

本発明の同一文書中に同一メディアが複数存在する場合における類似度設定の方法を説明するための図



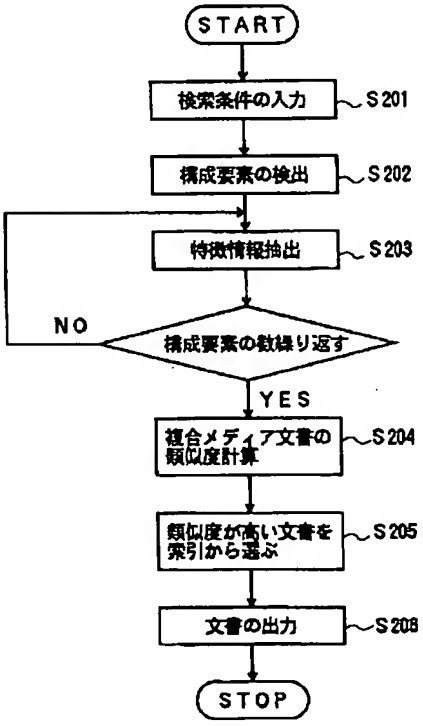

[Drawing 5]

本発明の複合メディア文書蓄積フェーズのフローチャート



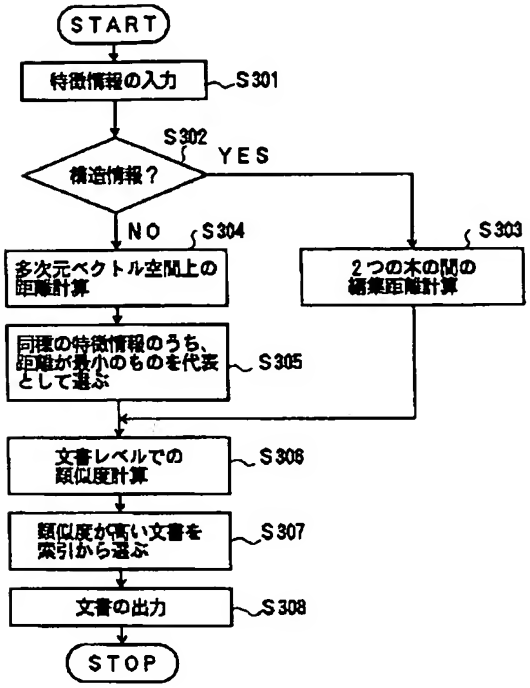

[Drawing 6]

本発明の複合メディア文書検索フェーズのフローチャート



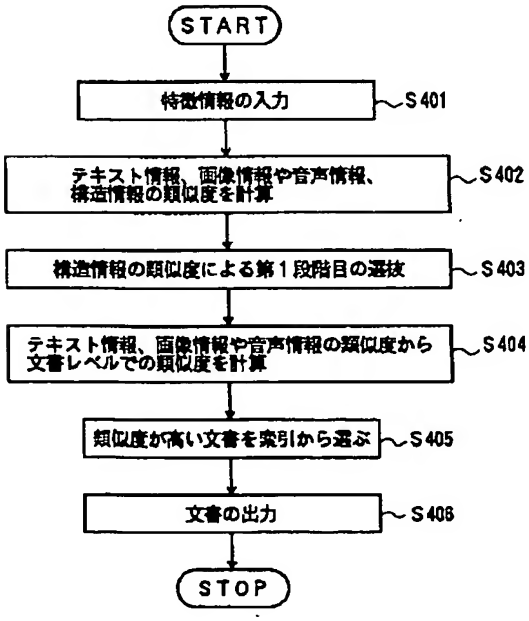

[Drawing 7]

本発明の類似度を求めるための文書比較を
行う際のフローチャート（その１）



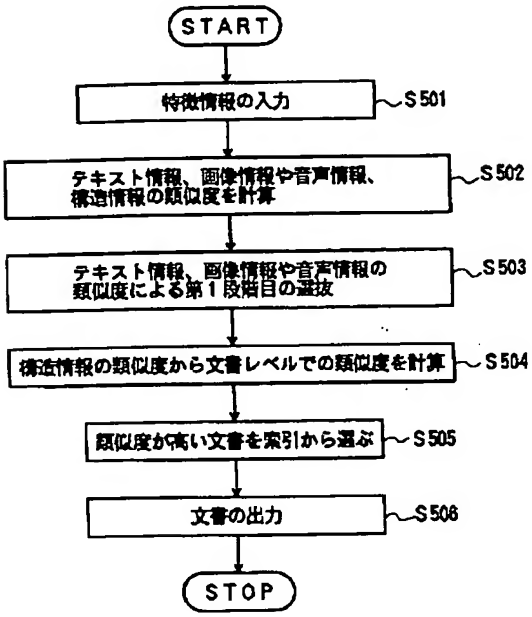

[Drawing 9]

本発明の類似度を求めるための文書比較を
行う際のフローチャート（その２）



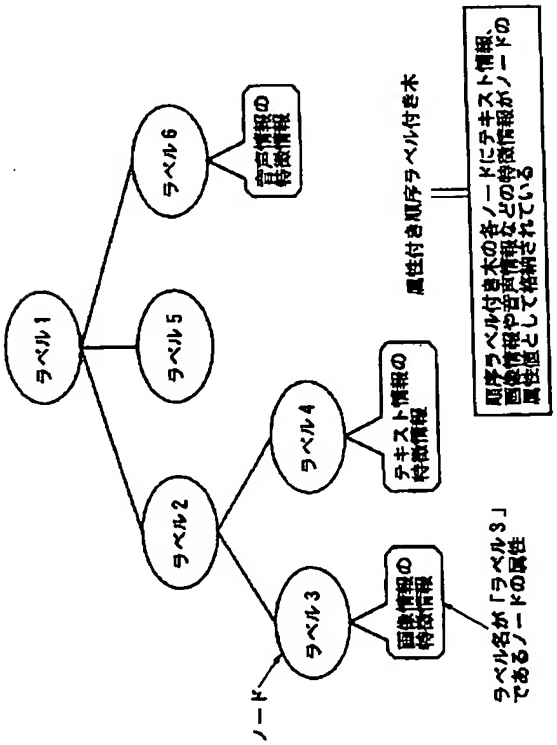

[Drawing 10]

本発明の類似度を求めるための文書比較を
行う際のフローチャート（その３）




[Drawing 11]

本発明の複合メディア文書を属性付き順序ラベル付き木として
表現することを説明するための図



[Translation done.]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（A）　(11)特許出願公開番号
特開2000－148793
（P2000－148793A）
(43)公開日　平成12年５月30日(2000.5.30)

| (51)Int.Cl.' | 識別記号 | FI | | テーマコード'（参考） |
|---|---|---|---|---|
| G06F 17/30 | | G06F 15/40 | 370G | 5B075 |
| | | 15/403 | 350C | |

審査請求　未請求　請求項の数63　ＯＬ　（全 21 頁）

(21)出願番号　特願平11－55950

(22)出願日　平成11年３月３日(1999.3.3)

(31)優先権主張番号　特願平10－258763
(32)優先日　平成10年９月11日(1998.9.11)
(33)優先権主張国　日本（ＪＰ）

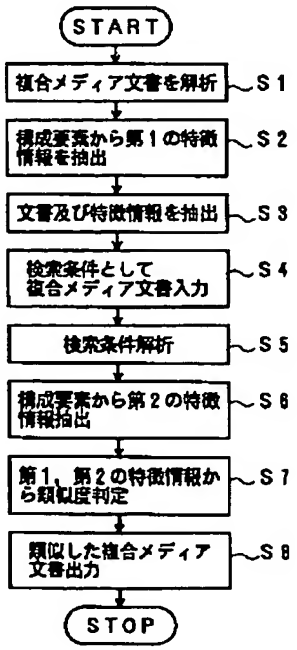
(71)出願人　000004226
日本電信電話株式会社
東京都千代田区大手町二丁目３番１号
(72)発明者　長谷川　知洋
東京都新宿区西新宿三丁目19番２号　日本電信電話株式会社内
(72)発明者　梅田　昌義
東京都新宿区西新宿三丁目19番２号　日本電信電話株式会社内
(74)代理人　100070150
弁理士　伊東　忠彦

最終頁に続く

(54)【発明の名称】　複合メディア文書の類似検索方法及び装置及び複合メディア文書の類似検索プログラムを格納した記憶媒体

(57)【要約】
【課題】　テキスト情報の類似性を重視した類似検索や画像情報の類似性を重視した類似検索や構造情報の類似性を重視した類似検索などが可能な複合メディア文書の類似検索方法及び装置及び複合メディア文書の類似検索プログラムを格納した記憶媒体を提供する。
【解決手段】　本発明は、検索条件として複合メディア文書が例示されると、例示された文書に含まれるテキスト情報、画像や音声の情報、構造情報などの特徴情報と蓄積された文書に含まれるテキスト情報、画像や音声の情報、構造情報などの特徴情報とをそれぞれ比較し、テキスト情報の類似度、画像情報の類似度、音声情報の類似度、構造情報の類似度を個々に計算し、それらに重みの値を掛け合わせ、総合的な評価値を計算したものを文書レベルでの類似度とし、重みの値を調節して、類似度が高い順に文書を検索結果として出力する。

本発明の原理を説明するための図



1

【特許請求の範囲】

【請求項１】　テキスト情報、画像情報、人間の声のデータである音声データに加え、ＣＤやレコードを含む楽曲データや音楽データを含む音声情報から構成される構造化文書である複合メディア文書の類似検索方法において、

与えられた前記複合メディア文書を構文解析し、

解析された結果得られる前記複合メディア文書の構成要素から第１の特徴情報を抽出し、

前記複合メディア文書及び抽出した前記特徴情報を蓄積し、

複合メディア文書を検索条件として入力し、

入力された前記検索条件を構文解析し、

解析された結果得られる前記検索条件の構成要素から第２の特徴情報を抽出し、

蓄積されている前記第１の特徴情報と前記第２の特徴情報に基づいて２つの複合メディア文書の類似度を判定し、類似した複合メディア文書を出力することを特徴とする複合メディア文書の類似検索方法。

【請求項２】　検索時において、

利用者が例示した文書を検索キーとして入力し、

例示された文書から前記第２の特徴情報を抽出し、

抽出された前記第２の特徴情報と前記第１の特徴情報により文書間の類似度を計算する請求項１記載の複合メディア文書の類似検索方法。

【請求項３】　前記複合メディア文書の類似度を計算する際に、

前記検索キーとして例示された前記文書を構成しているテキスト、画像、音声を含むメディアの情報及び構造情報と、

蓄積されている前記文書を構成しているテキスト、画像、音声を含むメディアの情報及び構造情報との構成要素毎の類似性判定結果に基づく評価値を、複合メディア文書全体類似度として設定する請求項１及び２記載の複合メディア文書の類似検索方法。

【請求項４】　前記構成要素毎の類似性判定結果として、

前記例示された文書のテキスト情報と、蓄積された前記文書のテキスト情報の類似性判定結果と、前記例示された文書の画像情報と、該蓄積された文書の画像情報との類似性判定結果と、該例示された文書の音声情報と、該蓄積された文書の音声情報との類似性判定結果と、該例示された文書の構造情報と、該蓄積された文書の構造情報との類似性判定結果を用いる請求項３記載の複合メディア文書の類似検索方法。

【請求項５】　前記複合メディア文書の類似度を計算する際に、

前記文書に含まれるテキスト情報、画像情報や音声情報、及び構造情報の特徴情報毎に、類似度を計算し、

前記類似度に重みの値を掛け、線形和をとったものを、

2

前記複合メディア文書全体としての類似度とする請求項３記載の複合メディア文書の類似検索方法。

【請求項６】　前記複合メディア文書の類似度を計算する際に、

同一文書中に同一メディアが複数存在する場合に、

前記例示された文書中に含まれるテキスト情報、画像情報、音声情報を含むメディア毎の全ての検索キーについて、蓄積された文書中の該メディアにおける該検索キーに対する全ての類似度を計算し、

前記類似度が最も高いものを前記検索キーに対する代表の類似度とする請求項２及び３記載の複合メディア文書の類似検索方法。

【請求項７】　前記複合メディア文書の類似度を計算する際に、

前記検索キーとして、前記例示された文書の構造情報と前記蓄積された文書の構造情報のそれぞれを、順序ラベル付木（ordered labeled tree）として表現し、

前記例示された文書を表現した前記順序ラベル付木と、前記蓄積された文書を表現した順序ラベル付木との形状を比較することで、文書の構造情報の類似度を設定する請求項１記載の複合メディア文書の類似検索方法。

【請求項８】　前記構造情報の類似度を設定する際に、

前記文書構造を木と見做し、前記例示された文書を表す木から蓄積された前記文書を表す木へ変換するために必要なノードの挿入、ノードの削除、ノード名の変更を含む編集操作を行った回数と、該編集操作を行うのに必要なコストから算出する編集距離を前記文書の類似度として設定する請求項７記載の複合メディア文書の類似検索方法。

【請求項９】　複合メディア文書の類似度を計算する際に、

文書に含まれるテキスト情報、画像情報や音声情報、構造情報の特徴情報に基づく類似度を計算し、

構造情報の特徴情報に基づく類似度計算結果に基づく第１段階目の選択を行い、

前記文書に含まれるテキスト情報、画像情報や音声情報の特徴情報に基づく類似度を、複合メディア文書全体としての類似度とする請求項３記載の複合メディア文書の類似検索方法。

【請求項１０】　複合メディア文書の類似度を計算する際に、

文書に含まれるテキスト情報、画像情報や音声情報、構造情報の特徴情報に基づく類似度を計算し、

前記テキスト情報、前記画像情報や前記音声情報の特徴情報に基づく類似度計算結果に基づく第１段階目の選択を行い、

前記文書に含まれる構造情報の特徴情報に基づく類似度を、複合メディア文書全体としての類似度とする請求項３記載の複合メディア文書の類似検索方法。

【請求項１１】　複合メディア文書の類似度を計算する

際に、
検索キーとして例示された文書中に同一メディアが複数存在する場合に、該メディアの文書レベルでの類似度を設定する請求項２及び３記載の複合メディア文書の類似検索方法。
【請求項１２】　前記メディアの文書レベルでの類似度を設定する際に、
検索キーとして例示された文書中に複数存在する前記メディアの各検索キーについて、
前記例示された文書中に含まれるテキスト情報、画像情報、音声情報を含むメディア毎の全ての検索キーについて、蓄積された文書中の該メディアにおける該検索キーに対する全ての類似度を計算し、
前記類似度が最も高いものを前記検索キーに対する代表の類似度とし、
前記代表の類似度の平均値を計算し、
前記メディアの文書レベルでの類似度を設定する請求項１１記載の複合メディア文書の類似検索方法。
【請求項１３】　前記メディアの文書レベルでの類似度を設定する際に、
検索キーとして例示された文書中に複数存在する前記メディアの各検索キーについて、
前記例示された文書中に含まれるテキスト情報、画像情報、音声情報を含むメディア毎の全ての検索キーについて、蓄積された文書中の該メディアにおける該検索キーに対する全ての類似度を計算し、
前記類似度が最も高いものを前記検索キーに対する代表の類似度とし、
前記代表の類似度のうち、最も類似度が高いものを前記メディアの文書レベルでの類似度とする請求項１１記載の複合メディア文書の類似検索方法。
【請求項１４】　複合メディア文書の類似検索を行う際に、
検索キーとして例示された文書の構造情報と蓄積された文書の構造情報のそれぞれを順序ラベル付き木（ordered labeled tree）として表現し、
それぞれの文書中の各メディアの特徴情報を前記順序ラベル付き木におけるノードの属性として格納した属性付き順序ラベル付き木として表現し、
前記例示された文書を表現した属性付き順序ラベル付き木と前記蓄積された文書を表現した属性付き順序ラベル付き木との属性と形状を比較することで、複合メディア文書の類似度を設定する請求項７記載の複合メディア文書の類似検索方法。
【請求項１５】　複合メディア文書の類似検索を行う際に、
前記例示された文書を表現した属性付き順序ラベル付き木の各ノードの属性である特徴情報と類似した特徴情報を属性として持つノードを持つ蓄積された文書を表現した属性付き順序ラベル付き木について、
前記ノードの構造的な位置関係の差異から複合メディア文書の類似度を設定する請求項１４記載の複合メディア文書の類似検索方法。
【請求項１６】　前記構造情報の類似度を設定する際に、
文書構造を順序ラベル付き木と見做し、該順序ラベル付き木に関する特徴情報に基づいて、多次元ベクトル空間上に該特徴情報を数値化してマッピングし、
前記ベクトル空間上での距離を文書の類似度として設定する請求項７記載の複合メディア文書の類似検索方法。
【請求項１７】　前記順序付きラベル木に関する特徴情報として、
前記順序ラベル付き木の各ノードの名前やノード数や各ノードの位置情報を数値化して利用することで文書の類似度を計算する請求項１６記載の複合メディア文書の類似検索方法。
【請求項１８】　前記特徴情報として、
テキスト情報であれば、テキストの記述内容が表す概念や各単語の出現頻度、画像情報であれば、画像の色相や彩度や輝度、色配置、音声情報であれば、音の強弱やメロディ、構造情報であれば、文書構造を順序ラベル付き木で表現した場合の木の形状やノードのラベル名、リンク情報等を、前記複合メディア文書の構成要素から抽出される特徴情報とする請求項１記載の複合メディア文書の類似検索方法。
【請求項１９】　前記類似度を判定する際に、
前記例示された文書に対する蓄積された文書の類似度を、
前記検索キーとして例示された前記文書を構成しているテキスト、画像、音声を含むメディアの情報及び構造情報と、蓄積されている前記文書を構成しているテキスト、画像、音声を含むメディアの情報及び構造情報との構成要素毎の類似性判定結果に基づく評価値を、複合メディア文書全体類似度として設定し、
前記蓄積された文書の類似度を降順に並べることで順位付けし、類似度を判定する請求項１記載の複合メディア文書の類似検索方法。
【請求項２０】　前記類似度を設定する際に、
前記複合メディア文書の各構成要素毎に、類似度を設定し、
前記文書に含まれるテキスト情報、画像情報や音声情報、及び構造情報の特徴情報毎に、類似度を計算し、
前記類似度に重みの値を掛け、線形和をとったものを、前記複合メディア文書全体としての類似度とする請求項３記載の複合メディア文書の類似検索方法。
【請求項２１】　前記複合メディア文書全体の類似度を設定する際に、
前記文書の構成要素毎の類似性判定結果に基づく評価値として、各構成要素の類似度そのもの、または、各構成要素の類似度に利用者から与えられた重みを掛け合わせ

たものを利用する請求項３記載の複合メディア文書の類似検索方法。

【請求項２２】　テキスト情報、画像情報、人間の声のデータである音声データに加え、ＣＤやレコードを含む楽曲データや音楽データを含む音声情報から構成される構造化文書である複合メディア文書の類似検索装置であって、

複合メディア文書を入力する複合メディア文書入力手段と、

前記複合メディア文書入力手段により与えられた前記複合メディア文書及び、入力された検索条件を構文解析する文書解析手段と、

前記文書解析手段で解析された結果得られる文書の構成要素から特徴情報を抽出する特徴情報抽出手段と、

前記複合メディア文書及び前記特徴情報抽出手段で抽出された前記特徴情報を蓄積する蓄積手段と、

複合メディア文書を検索条件として入力する検索条件入力手段と、

前記蓄積手段に蓄積されている前記複合メディア文書の特徴情報と、入力された前記検索条件を前記文書解析手段で解析した結果に基づいて前記特徴情報抽出手段で抽出された特徴情報に基づいて２つの複合メディア文書の類似度を判定する文書比較手段と、

前記文書比較手段で判定された類似度に基づいて、類似した複合メディア文書を出力する出力手段とを有することを特徴とする複合メディア文書の類似検索装置。

【請求項２３】　前記検索条件入力手段は、

利用者が例示した文書を検索キーとして入力する手段を含み、

前記特徴情報抽出手段は、

与えられた前記複合メディア文書から検索キーとして例示された複合メディア文書を抽出する入力文書特徴情報抽出手段と、

前記利用者から例示された文書から前記検索情報の特徴情報を抽出する検索特徴情報抽出手段を含み、

前記文書比較手段は、

前記入力文書特徴情報抽出手段で抽出された入力文書特徴情報と、前記検索特徴情報抽出手段で抽出された検索特徴情報により、前記複合メディア文書と前記検索キーとして例示された複合メディア文書間の類似度を計算する類似度計算手段を含む請求項２２記載の複合メディア文書の類似検索装置。

【請求項２４】　前記類似度計算手段は、

前記検索条件入力手段で前記検索キーとして前記例示された文書を構成しているテキスト情報、画像情報、音声情報及び構造情報と、前記蓄積手段に蓄積されている前記複合メディア文書を構成しているテキスト情報、画像情報、音声情報及び構造情報との構成要素毎の類似性判定結果に基づく評価値を、複合メディア文書全体の類似度として設定する類似度設定手段を含む請求項２１及び２２記載の複合メディア文書の類似検索装置。

【請求項２５】　前記類似度設定手段は、

前記構成要素毎の類似性判定結果として、

前記例示された文書のテキスト情報と前記蓄積手段に蓄積されている前記文書のテキスト情報の類似性判定結果と、該例示された文書の画像情報と、該蓄積手段に蓄積されている文書の画像情報との類似性判定結果と、該例示された文書の音声情報と該蓄積手段に蓄積されている文書の音声情報との類似性判定結果と、該例示さた文書の構造情報と該蓄積手段に蓄積されている文書の構造情報との類似性判定結果を用いる請求項２４記載の複合メディア文書の類似検索装置。

【請求項２６】　前類似度設定手段は、

前記文書に含まれるテキスト情報、画像情報や音声情報、及び構造情報の特徴情報毎に、類似度を計算し、該類似度に重みの値を掛け、線形和をとったものを、前記複合メディア文書全体としての類似度とする線形和算出手段を含む請求項２４記載の複合メディア文書の類似検索装置。

【請求項２７】　前記文書比較手段は、

同一文書中に同一メディアが複数存在する場合に、前記例示された文書中に含まれるテキスト、画像、音声を含むメディア毎の全ての検索キーについて、蓄積された文書中のメディアにおける該検索キーに対する全ての類似度を計算し、該類似度が最も高いものを前記検索キーに対する代表の類似度とする代表類似度決定手段を含む請求項２２及び２３記載の複合メディア文書の類似検索装置。

【請求項２８】　前記類似度計算手段は、

前記検索キーとして、前記例示された文書の構造情報と蓄積された文書の構造情報のそれぞれを、順序ラベル付木（ordered labeled tree）として表現し、

前記例示された文書を表現した前記順序ラベル付木と、前記蓄積された文書を表現した順序ラベル付木との形状を比較することで、文書の構造情報の類似度を設定する順序ラベル付木形状比較手段を含む請求項２３記載の複合メディア文書の類似検索装置。

【請求項２９】　前記順序ラベル付木形状比較手段は、

前記構造情報の類似度を判定する際に、前記文書構造を木と見做し、前記例示された文書を表す木から蓄積された前記文書を表す木へ変換するために必要なノードの挿入、ノードの削除、ノード名の変更を含む編集操作を行った回数と、該編集操作を行うのに必要なコストから算出する編集距離を前記文書の類似度として設定する編集距離算出手段を含む請求項２８記載の複合メディア文書の類似検索装置。

【請求項３０】　前記類似度計算手段は、

文書に含まれるテキスト情報、画像情報や音声情報、構造情報の特徴情報に基づく類似度を計算する手段と、

構造情報の特徴情報に基づく類似度計算結果に基づく第

１段階目の選択を行う手段と、
前記文書に含まれるテキスト情報、画像情報や音声情報の特徴情報に基づく類似度を、複合メディア文書全体としての類似度とする手段とを含む請求項２４記載の複合メディア文書の類似検索装置。
【請求項３１】　前記類似度計算手段は、
文書に含まれるテキスト情報、画像情報や音声情報、構造情報の特徴情報に基づく類似度を計算する手段と、
前記テキスト情報、前記画像情報や前記音声情報の特徴情報に基づく類似度計算結果に基づく第１段階目の選択を行う手段と、
前記文書に含まれる構造情報の特徴情報に基づく類似度を、複合メディア文書全体としての類似度とする手段とを含む請求項２４記載の複合メディア文書の類似検索装置。
【請求項３２】　前記類似度計算手段は、
検索キーとして例示された文書中に同一メディアが複数存在する場合に、該メディアの文書レベルでの類似度を設定する文書レベル類似度計算手段を含む請求項２３及び２４記載の複合メディア文書の類似検索装置。
【請求項３３】　前記文書レベル類似度計算手段は、
検索キーとして例示された文書中に複数存在する前記メディアの各検索キーについて、前記例示された文書中に含まれるテキスト情報、画像情報、音声情報を含むメディア毎の全ての検索キーについて、蓄積された文書中の該メディアにおける該検索キーに対する全ての類似度を計算する手段と、
前記類似度が最も高いものを前記検索キーに対する代表の類似度とする手段と、
前記代表の類似度の平均値を計算する手段と、
前記メディアの文書レベルでの類似度を設定する文書レベル類似度設定手段とを含む請求項３２記載の複合メディア文書の類似検索装置。
【請求項３４】　前記文書レベル類似度設定手段は、
検索キーとして例示された文書中に複数存在する前記メディアの各検索キーについて、該例示された文書中に含まれるテキスト情報、画像情報、音声情報を含むメディア毎の全ての検索キーについて、蓄積された文書中の該メディアにおける該検索キーに対する全ての類似度を計算する手段と、
前記類似度が最も高いものを前記検索キーに対する代表の類似度とする手段と、
前記代表の類似度のうち、最も類似度が高いものを前記メディアの文書レベルでの類似度とする手段とを含む請求項３２記載の複合メディア文書の類似検索装置。
【請求項３５】　前記類似度計算手段は、
検索キーとして例示された文書の構造情報と蓄積された文書の構造情報のそれぞれを順序ラベル付き木（ordered labeled tree）として表現し、それぞれの文書中の各メディアの特徴情報を前記順序ラベル付き木におけるノードの属性として格納した属性付き順序ラベル付き木として表現し、前記例示された文書を表現した属性付き順序ラベル付き木と前記蓄積された文書を表現した属性付き順序ラベル付き木との属性と形状を比較することで、複合文書の類似度を設定する類似検索手段を含む請求項２８記載の複合メディア文書の類似検索装置。
【請求項３６】　前記類似検索手段は、
例示された文書を表現した属性付き順序ラベル付き木の各ノードの属性である特徴情報と類似した特徴情報を属性として持つノードを持つ蓄積された文書を表現した属性付き順序ラベル付き木について、該ノードの構造的な位置関係の差異から複合メディア文書の類似度を設定する手段を含む請求項３５記載の複合メディア文書の類似検索装置。
【請求項３７】　前記順序ラベル付き木形状比較手段は、
文書構造を順序ラベル付き木と見做し、該順序ラベル付き木に関する特徴情報に基づいて、多次元ベクトル空間上に該特徴情報を数値化してマッピングする手段と、
前記ベクトル空間上での距離を文書の類似度として設定する手段とを含む請求項３０記載の複合メディア文書の類似検索装置。
【請求項３８】　前記順序付きラベル木に関する特徴情報として、
前記順序ラベル付き木の各ノードの名前やノード数や各ノードの位置情報を数値化して利用することで文書の類似度を計算する請求項３７記載の複合メディア文書の類似検索装置。
【請求項３９】　前記特徴情報として、
テキスト情報であれば、テキストの記述内容が表す概念や各単語の出現頻度、画像情報であれば、画像の色相や彩度や輝度、色配置、音声情報であれば、音の強弱やメロディ、構造情報であれば、文書構造を順序ラベル付き木で表現した場合の木の形状やノードのラベル名、リンク情報などを、前記複合メディア文書の構成要素から抽出される特徴情報とする請求項２２記載の複合メディア文書の類似検索装置。
【請求項４０】　前記文書比較手段は、
前記例示された文書に対する蓄積された文書の類似度を、
前記検索キーとして例示された前記文書を構成しているテキスト、画像、音声を含むメディアの情報及び構造情報と、蓄積されている前記文書を構成しているテキスト、画像、音声を含むメディアの情報及び構造情報との構成要素毎の類似性判定結果に基づく評価値を、複合メディア文書全体類似度として設定する手段と、
前記蓄積された文書の類似度を降順に並べることで順位付けし、類似度を判定する手段とを含む請求項２２記載の複合メディア文書の類似検索装置。
【請求項４１】　前記類似度設定手段は、

前記複合メディア文書の各構成要素毎に、類似度を設定する手段と、
前記文書に含まれるテキスト情報、画像情報や音声情報、及び構造情報の特徴情報毎に、類似度を計算する手段と、
前記類似度に重みの値を掛け、線形和をとったものを、前記複合メディア文書全体としての類似度とする手段とを含む請求項２４記載の複合メディア文書の類似検索装置。

【請求項４２】　前記類似度設定手段は、
前記文書の構成要素毎の類似性判定結果に基づく評価値として、各構成要素の類似度そのもの、または、各構成要素の類似度に利用者から与えられた重みを掛け合わせたものを利用する手段を含む請求項２４記載の複合メディア文書の類似検索装置。

【請求項４３】　テキスト情報、画像情報、人間の声のデータである音声データに加え、ＣＤやレコードを含む楽曲データや音楽データを含む音声情報から構成される構造化文書である複合メディア文書の類似検索プログラムを格納した記憶媒体であって、
与えられた前記複合メディア文書及び、入力された検索条件を構文解析する文書解析プロセスと、
前記文書解析プロセスで解析された結果、得られる文書の構成要素から特徴情報を抽出する特徴情報抽出プロセスと、
前記複合メディア文書及び前記特徴情報抽出プロセスで抽出された前記特徴情報を記憶手段に格納する格納プロセスと、
複合メディア文書を検索条件として入力させる検索条件入力プロセスと、
前記記憶手段に蓄積されている前記複合メディア文書の特徴情報と、入力された前記検索条件を前記文書解析プロセスで解析した結果に基づいて前記特徴情報抽出プロセスで抽出された特徴情報から２つの複合メディア文書の類似度を判定する文書比較プロセスと、
前記文書比較プロセスで判定された類似度に基づいて、類似した複合メディア文書を出力させる出力プロセスとを有することを特徴とする複合メディア文書の類似検索プログラムを格納した記憶媒体。

【請求項４４】　前記検索条件入力プロセスは、
利用者が例示した文書を検索キーとして入力するプロセスを含み、
前記特徴情報抽出プロセスは、
与えられた前記複合メディア文書から特徴情報を抽出する入力文書特徴情報抽出プロセスと、
前記利用者から例示された文書から検索キーとして例示された複合メディア文書の特徴情報を抽出する検索特徴情報抽出プロセスを含み、
前記文書比較プロセスは、
前記入力文書特徴情報抽出プロセスで抽出された入力文書特徴情報と、前記検索特徴情報抽出プロセスで抽出された検索特徴情報により、前記複合メディア文書と前記検索キーとして例示された複合メディア文書間の類似度を計算する類似度計算プロセスを含む請求項４３記載の複合メディア文書の類似検索プログラムを格納した記憶媒体。

【請求項４５】　前記類似度計算プロセスは、
前記検索条件入力プロセスで前記検索キーとして例示された文書を構成しているテキスト情報、画像情報、音声情報及び構造情報と、前記記憶手段に蓄積されている前記複合メディア文書を構成しているテキスト情報、画像情報、音声情報及び構造情報との構成要素毎の類似性判定結果に基づく評価値を、複合メディア文書全体の類似度として設定する類似度設定プロセスを含む請求項４３及び４４記載の複合メディア文書の類似検索プログラムを格納した記憶媒体。

【請求項４６】　前記類似度設定プロセスは、
前記構成要素毎の類似性判定結果として、
前記例示された文書のテキスト情報と前記記憶手段に蓄積されている前記文書のテキスト情報の類似性判定結果と、該例示された文書の画像情報と、該記憶手段に蓄積されている文書の画像情報との類似性判定結果と、該例示された文書の音声情報と該記憶手段に蓄積されている文書の音声情報との類似性判定結果と、該例示さた文書の構造情報と該記憶手段に蓄積されている文書の構造情報との類似性判定結果を用いる請求項４５記載の複合メディア文書の類似検索プログラムを格納した記憶媒体。

【請求項４７】　前類似度設定プロセスは、
前記文書に含まれるテキスト情報、画像情報や音声情報、及び構造情報の特徴情報毎に、類似度を計算し、該類似度に重みの値を掛け、線形和をとったものを、前記複合メディア文書全体としての類似度とする線形和算出プロセスを含む請求項４５記載の複合メディア文書の類似検索プログラムを格納した記憶媒体。

【請求項４８】　前記文書比較プロセスは、
同一文書中に同一メディアが複数存在する場合に、前記例示された文書中に含まれるテキスト、画像、音声を含むメディア毎の全ての検索キーについて、蓄積された文書中のメディアにおける該検索キーに対する全ての類似度を計算し、該類似度が最も高いものを前記検索キーに対する代表の類似度とする代表類似度決定プロセスを含む請求項４３及び４４記載の複合メディア文書の類似検索プログラムを格納した記憶媒体。

【請求項４９】　前記類似度計算プロセスは、
前記検索キーとして、前記例示された文書の構造情報と蓄積された文書の構造情報のそれぞれを、順序ラベル付木（ordered labeled tree）として表現し、前記例示された文書を表現した前記順序ラベル付木と、前記蓄積された文書を表現した順序ラベル付木との形状を比較することで、文書の構造情報の類似度を設定する順序ラベル

付木形状比較プロセスを含む請求項４４記載の複合メディア文書の類似検索プログラムを格納した記憶媒体。

【請求項５０】　前記順序ラベル付木形状比較プロセスは、

前記構造情報の類似度を判定する際に、前記文書構造を木と見做し、前記例示された文書を表す木から蓄積された前記文書を表す木へ変換するために必要なノードの挿入、ノードの削除、ノード名の変更を含む編集操作を行った回数と、該編集操作を行うのに必要なコストから算出する編集距離を前記文書の類似度として設定する編集距離算出プロセスを含む請求項４９記載の複合メディア文書の類似検索プログラムを格納した記憶媒体。

【請求項５１】　前記類似度計算プロセスは、

文書に含まれるテキスト情報、画像情報や音声情報、構造情報の特徴情報に基づく類似度を計算するプロセスと、

構造情報の特徴情報に基づく類似度計算結果に基づく第１段階目の選択を行うプロセスと、

前記文書に含まれるテキスト情報、画像情報や音声情報の特徴情報に基づく類似度を、複合メディア文書全体としての類似度とするプロセスとを含む請求項４４記載の複合メディア文書の類似検索プログラムを格納した記憶媒体。

【請求項５２】　前記類似度計算プロセスは、

文書に含まれるテキスト情報、画像情報や音声情報、構造情報の特徴情報に基づく類似度を計算するプロセスと、

前記テキスト情報、前記画像情報や前記音声情報の特徴情報に基づく類似度計算結果に基づく第１段階目の選択を行うプロセスと、

前記文書に含まれる構造情報の特徴情報に基づく類似度を、複合メディア文書全体としての類似度とするプロセスとを含む請求項４３及び、４４記載の複合メディア文書の類似検索プログラムを格納した記憶媒体。

【請求項５３】　前記類似度計算プロセスは、

検索キーとして例示された文書中に同一メディアが複数存在する場合に、該メディアの文書レベルでの類似度を設定する文書レベル類似度計算プロセスを含む請求項４３及び４４記載の複合メディア文書の類似検索プログラムを格納した記憶媒体。

【請求項５４】　前記文書レベル類似度計算プロセスは、

検索キーとして例示された文書中に複数存在する前記メディアの各検索キーについて、前記例示された文書中に含まれるテキスト情報、画像情報、音声情報を含むメディア毎の全ての検索キーについて、蓄積された文書中の該メディアにおける該検索キーに対する全ての類似度を計算するプロセスと、

前記類似度が最も高いものを前記検索キーに対する代表の類似度とするプロセスと、

前記代表の類似度の平均値を計算するプロセスと、

前記メディアの文書レベルでの類似度を設定する文書レベル類似度設定プロセスとを含む請求項５３記載の複合メディア文書の類似検索プログラムを格納した記憶媒体。

【請求項５５】　前記文書レベル類似度設定プロセスは、

検索キーとして例示された文書中に複数存在する前記メディアの各検索キーについて、該例示された文書中に含まれるテキスト情報、画像情報、音声情報を含むメディア毎の全ての検索キーについて、蓄積された文書中の該メディアにおける該検索キーに対する全ての類似度を計算するプロセスと、

前記類似度が最も高いものを前記検索キーに対する代表の類似度とするプロセスと、

前記代表の類似度のうち、最も類似度が高いものを前記メディアの文書レベルでの類似度とするプロセスとを含む請求項５４記載の複合メディア文書の類似検索プログラムを格納した記憶媒体。

【請求項５６】　前記類似度計算プロセスは、

検索キーとして例示された文書の構造情報と蓄積された文書の構造情報のそれぞれを順序ラベル付き木（ordered labeled tree）として表現し、それぞれの文書中の各メディアの特徴情報を前記順序ラベル付き木におけるノードの属性として格納した属性付き順序ラベル付き木として表現し、前記例示された文書を表現した属性付き順序ラベル付き木と前記蓄積された文書を表現した属性付き順序ラベル付き木との属性と形状を比較することで、複合文書の類似度を設定する類似検索プロセスを含む請求項４９記載の複合メディア文書の類似検索プログラムを格納した記憶媒体。

【請求項５７】　前記類似検索プロセスは、

例示された文書を表現した属性付き順序ラベル付き木の各ノードの属性である特徴情報と類似した特徴情報を属性として持つノードを持つ蓄積された文書を表現した属性付き順序ラベル付き木について、該ノードの構造的な位置関係の差異から複合メディア文書の類似度を設定するプロセスを含む請求項５６記載の複合メディア文書の類似検索プログラムを格納した記憶媒体。

【請求項５８】　前記順序ラベル付き木形状比較プロセスは、

文書構造を順序ラベル付き木と見做し、該順序ラベル付き木に関する特徴情報に基づいて、多次元ベクトル空間上に該特徴情報を数値化してマッピングするプロセスと、

前記ベクトル空間上での距離を文書の類似度として設定するプロセスとを含む請求項４９記載の複合メディア文書の類似検索プログラムを格納した記憶媒体。

【請求項５９】　前記順序付きラベル木に関する特徴情報として、

前記順序ラベル付き木の各ノードの名前やノード数や各ノードの位置情報を数値化して利用することで文書の類似度を計算する請求項５８記載の複合メディア文書の類似検索プログラムを格納した記憶媒体。

【請求項６０】　前記特徴情報として、

テキスト情報であれば、テキストの記述内容が表す概念や各単語の出現頻度、画像情報であれば、画像の色相や彩度や輝度、色配置、音声情報であれば、音の強弱やメロディ、構造情報であれば、文書構造を順序ラベル付き木で表現した場合の木の形状やノードのラベル名、リンク情報などを、前記複合メディア文書の構成要素から抽出される特徴情報とする請求項４３記載の複合メディア文書の類似検索プログラムを格納した記憶媒体。

【請求項６１】　前記文書比較プロセスは、

前記例示された文書に対する蓄積された文書の類似度を、

前記検索キーとして例示された前記文書を構成しているテキスト、画像、音声を含むメディアの情報及び構造情報と、蓄積されている前記文書を構成しているテキスト、画像、音声を含むメディアの情報及び構造情報との構成要素毎の類似性判定結果に基づく評価値を、複合メディア文書全体類似度として設定するプロセスと、

前記蓄積された文書の類似度を降順に並べることで順位付けし、類似度を判定するプロセスとを含む請求項４３記載の複合メディア文書の類似検索プログラムを格納した記憶媒体。

【請求項６２】　前記類似度設定プロセスは、

前記複合メディア文書の各構成要素毎に、類似度を設定するプロセスと、

前記文書に含まれるテキスト情報、画像情報や音声情報、及び構造情報の特徴情報毎に、類似度を計算するプロセスと、

前記類似度に重みの値を掛け、線形和をとったものを、前記複合メディア文書全体としての類似度とするプロセスとを含む請求項４４記載の複合メディア文書の類似検索プログラムを格納した記憶媒体。

【請求項６３】　前記類似度設定プロセスは、

前記文書の構成要素毎の類似性判定結果に基づく評価値として、各構成要素の類似度そのもの、または、各構成要素の類似度に利用者から与えられた重みを掛け合わせたものを利用するプロセスを含む請求項４４記載の複合メディア文書の類似検索プログラムを格納した記憶媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、複合メディア文書の類似検索方法及び装置及び複合メディア文書の類似検索プログラムを格納した記憶媒体に係り、特に、複合メディア文書データベースについて、類似した複合メディア文書を検索するための複合メディア文書の類似検索方法及び装置及び複合メディア文書の類似検索プログラムを格納した記憶媒体に関する。

【０００２】

【従来の技術】構造化文書検索方法に関する従来の技術としては、奈良先端大学のStructureIndex とContent Index の組み合わせ（金本、加藤、絹谷、吉川、“効率的な更新が可能な構造化文書検索法”）等がある。これらのシステムでは、ＳＧＭＬ文書やＸＭＬ文書などの構造化文書に格納されたテキスト情報や文書構造の情報（エレメント名など）の出現位置に関する転置ファイルなどの索引を予め用意し、テキスト情報または、テキスト情報と構造情報の組み合わせを検索キーとして与え、与えたキーワードが文書に含まれているか否かを判定する論理検索や、指定した範囲内に与えたキーワードが出現するか否かを判定する範囲検索を行うことで、構造化文書の検索を可能にしている。

【０００３】ジャストシステムのConceptBase やコマツソフトのVextSearchなどのシステムで、検索キーとして自然文などで入力されたテキストに含まれる概念と検索対象となるテキストに含まれる概念とを比較して、類似文書（テキストのみ）の検索を可能にしている。

【０００４】

【発明が解決しようとする課題】しかしながら、上記の従来の方法を複合メディア文書を対象とした類似検索に適用しようとした場合、複合メディア文書を対象とした類似検索法が確立されていないため、以下のような問題が発生する。

・利用者が構造化文書中のエレメント名（注：構造情報を表すタグ名）などの文書構造に関する情報を予め知らなければ、検索キーの一部に構造情報を与えた検索を行うことができない。

【０００５】・画像や音声などテキスト以外のメディアの情報や構造情報を検索キーとして使用した複合メディア文書の類似検索ができない。

本発明は、上記の点に鑑みなされたもので、検索条件として複合メディア文書が例示されると、例示された文書に含まれるテキスト情報、画像や音声の情報、構造情報などの特徴情報と蓄積された文書に含まれるテキスト情報、画像や音声の情報、構造情報などの特徴情報とをそれぞれ比較し、テキスト情報の類似度、画像情報の類似度、音声情報の類似度、構造情報の類似度を個々に計算し、それらに重みの値を掛け合わせ、総合的な評価値を計算したものを文書レベルでの類似度とし、重みの値を調節することで、テキスト情報の類似性を重視した類似検索や画像情報の類似性を重視した類似検索や構造情報の類似性を重視した類似検索などが可能な複合メディア文書の類似検索方法及び装置及び複合メディア文書の類似検索プログラムを格納した記憶媒体を提供することを目的とする。

【０００６】

【課題を解決するための手段】図１は、本発明の原理を説明するための図である。本発明（請求項１）は、テキスト情報、画像情報、人間の声のデータである音声データに加え、ＣＤやレコードを含む楽曲データや音楽データを含む音声情報から構成される構造化文書である複合メディア文書の類似検索方法において、与えられた複合メディア文書を構文解析し（ステップ１）、解析された結果得られる複合メディア文書の構成要素から第１の特徴情報を抽出し（ステップ２）、文書及び抽出した特徴情報を蓄積し（ステップ３）、複合メディア文書を検索条件として入力し（ステップ４）、入力された検索条件を構文解析し（ステップ５）、解析された結果得られる検索条件の構成要素から第２の特徴情報を抽出し（ステップ６）、蓄積されている第１の特徴情報と第２の特徴情報に基づいて２つの複合メディア文書の類似度を判定し（ステップ７）、類似した複合メディア文書を出力する（ステップ８）。

【０００７】本発明（請求項２）は、検索時において、利用者が例示した文書を検索キーとして入力し、例示された文書から第２の特徴情報を抽出し、抽出された第２の特徴情報と第１の特徴情報により文書間の類似度を計算する。本発明（請求項３）は、複合メディア文書の類似度を計算する際に、検索キーとして例示された文書を構成しているテキスト、画像、音声を含むメディアの情報及び構造情報と、蓄積されている文書を構成しているテキスト、画像、音声を含むメディアの情報及び構造情報との構成要素毎の類似性判定結果に基づく評価値を、複合メディア文書全体の類似度として設定する。

【０００８】本発明（請求項４）は、構成要素毎の類似性判定結果として、例示された文書のテキスト情報と、蓄積された文書のテキスト情報の類似性判定結果と、例示された文書の画像情報と、該蓄積された文書の画像情報との類似性判定結果と、該例示された文書の音声情報と、該蓄積された文書の音声情報との類似性判定結果と、該例示された文書の構造情報と、該蓄積された文書の構造情報との類似性判定結果を用いる。

【０００９】本発明（請求項５）は、複合メディア文書の類似度を計算する際に、文書に含まれるテキスト情報、画像情報や音声情報、及び構造情報の特徴情報毎に、類似度を計算し、類似度に重みの値を掛け、線形和をとったものを、複合メディア文書全体としての類似度とする。

【００１０】本発明（請求項６）は、複合メディア文書の類似度を計算する際に、同一文書中に同一メディアが複数存在する場合に、例示された文書中に含まれるテキスト情報、画像情報、音声情報を含むメディア毎の全ての検索キーについて、蓄積された文書中の該メディアにおける該検索キーに対する全ての類似度を計算し、類似度が最も高いものを検索キーに対する代表の類似度とする。

【００１１】本発明（請求項７）は、複合メディア文書の類似度を計算する際に、検索キーとして、例示された文書の構造情報と蓄積された文書の構造情報のそれぞれを、順序ラベル付木（ordered labeled tree）として表現し、例示された文書を表現した順序ラベル付木と、蓄積された文書を表現した順序ラベル付木との形状を比較することで、文書の構造情報の類似度を設定する。

【００１２】本発明（請求項８）は、構造情報の類似度を設定する際に、文書構造を木と見做し、例示された文書を表す木から蓄積された文書を表す木へ変換するために必要なノードの挿入、ノードの削除、ノード名の変更を含む編集操作を行った回数と、該編集操作を行うのに必要なコストから算出する編集距離を文書の類似度として設定する。

【００１３】本発明（請求項９）は、複合メディア文書の類似度を計算する際に、文書に含まれるテキスト情報、画像情報や音声情報、構造情報の特徴情報に基づく類似度を計算し、構造情報の特徴情報に基づく類似度計算結果に基づく第１段階目の選択を行い、文書に含まれるテキスト情報、画像情報や音声情報の特徴情報に基づく類似度を、複合メディア文書全体としての類似度とする。

【００１４】本発明（請求項１０）は、複合メディア文書の類似度を計算する際に、文書に含まれるテキスト情報、画像情報や音声情報、構造情報の特徴情報に基づく類似度を計算し、テキスト情報、画像情報や音声情報の特徴情報に基づく類似度計算結果に基づく第１段階目の選択を行い、文書に含まれる構造情報の特徴情報に基づく類似度を、複合メディア文書全体としての類似度とする。

【００１５】本発明（請求項１１）は、複合メディア文書の類似度を計算する際に、検索キーとして例示された文書中に同一メディアが複数存在する場合に、該メディアの文書レベルでの類似度を設定する。本発明（請求項１２）は、メディアの文書レベルでの類似度を設定する際に、検索キーとして例示された文書中に複数存在するメディアの各検索キーについて、例示された文書中に含まれるテキスト情報、画像情報、音声情報を含むメディア毎の全ての検索キーについて、蓄積された文書中の該メディアにおける該検索キーに対する全ての類似度を計算し、類似度が最も高いものを検索キーに対する代表の類似度とし、代表の類似度の平均値を計算し、メディアの文書レベルでの類似度を設定する。

【００１６】本発明（請求項１３）は、メディアの文書レベルでの類似度を設定する際に、検索キーとして例示された文書中に複数存在するメディアの各検索キーについて、例示された文書中に含まれるテキスト情報、画像情報、音声情報を含むメディア毎の全ての検索キーについて、蓄積された文書中の該メディアにおける該検索キーに対する全ての類似度を計算し、類似度が最も高いも

のを検索キーに対する代表の類似度とし、代表の類似度のうち、最も類似度が高いものをメディアの文書レベルでの類似度とする。

【００１７】本発明（請求項１４）は、複合メディア文書の類似検索を行う際に、検索キーとして例示された文書の構造情報と蓄積された文書の構造情報のそれぞれを順序ラベル付き木（ordered labeled tree）として表現し、それぞれの文書中の各メディアの特徴情報を順序ラベル付き木におけるノードの属性として格納した属性付き順序ラベル付き木として表現し、例示された文書を表現した属性付き順序ラベル付き木と蓄積された文書を表現した属性付き順序ラベル付き木との属性と形状を比較することで、複合メディア文書の類似度を設定する。

【００１８】本発明（請求項１５）は、複合メディア文書の類似検索を行う際に、例示された文書を表現した属性付き順序ラベル付き木の各ノードの属性である特徴情報と類似した特徴情報を属性として持つノードを持つ蓄積された文書を表現した属性付き順序ラベル付き木について、ノードの構造的な位置関係の差異から複合メディア文書の類似度を設定する。

【００１９】本発明（請求項１６）は、構造情報の類似度を設定する際に、文書構造を順序ラベル付き木と見做し、該順序ラベル付き木に関する特徴情報に基づいて、多次元ベクトル空間上に該特徴情報を数値化してマッピングし、ベクトル空間上での距離を文書の類似度として設定する。本発明（請求項１７）は、順序付きラベル木に関する特徴情報として、順序ラベル付き木の各ノードの名前やノード数や各ノードの位置情報を数値化して利用することで文書の類似度を計算する。

【００２０】本発明（請求項１８）は、特徴情報として、テキスト情報であれば、テキストの記述内容が表す概念や各単語の出現頻度、画像情報であれば、画像の色相や彩度や輝度、色配置、音声情報であれば、音の強弱やメロディ、構造情報であれば、文書構造を順序ラベル付き木で表現した場合の木の形状やノードのラベル名、リンク情報などを、複合メディア文書の構成要素から抽出される特徴情報とする。

【００２１】本発明（請求項１９）は、類似度を判定する際に、例示された文書に対する蓄積された文書の類似度を、検索キーとして例示された文書を構成しているテキスト、画像、音声を含むメディアの情報及び構造情報と、蓄積されている文書を構成しているテキスト、画像、音声を含むメディアの情報及び構造情報との構成要素毎の類似性判定結果に基づく評価値を、複合メディア文書全体類似度として設定し、蓄積された文書の類似度を降順に並べることで順位付けし、類似度を判定する。

【００２２】本発明（請求項２０）は、類似度を設定する際に、複合メディア文書の各構成要素毎に、類似度を設定し、文書に含まれるテキスト情報、画像情報や音声情報、及び構造情報の特徴情報毎に、類似度を計算し、類似度に重みの値を掛け、線形和をとったものを、複合メディア文書全体としての類似度とする。

【００２３】本発明（請求項２１）は、複合メディア文書全体の類似度を設定する際に、文書の構成要素毎の類似性判定結果に基づく評価値として、各構成要素の類似度そのもの、または、各構成要素の類似度に利用者から与えられた重みを掛け合わせたものを利用する。図２は、本発明の原理構成図である。

【００２４】本発明（請求項２２）は、テキスト情報、画像情報、人間の声のデータである音声データに加え、ＣＤやレコードを含む楽曲データや音楽データを含む音声情報から構成される構造化文書である複合メディア文書の類似検索装置であって、複合メディア文書を入力する複合メディア文書入力手段１０と、複合メディア文書入力手段１０により与えられた複合メディア文書及び、入力された検索条件を構文解析する文書解析手段４０と、文書解析手段４０で解析された結果得られる文書の構成要素から特徴情報を抽出する特徴情報抽出手段５０と、複合メディア文書及び特徴情報抽出手段５０で抽出された特徴情報を蓄積する蓄積手段６０と、複合メディア文書を検索条件として入力する検索条件入力手段３０と、蓄積手段６０に蓄積されている複合メディア文書の特徴情報と、入力された検索条件を文書解析手段４０で解析した結果に基づいて特徴情報抽出手段５０で抽出された特徴情報に基づいて２つの複合メディア文書の類似度を判定する文書比較手段８０と、文書比較手段８０で判定された類似度に基づいて、類似した複合メディア文書を出力する出力手段９０とを有する。

【００２５】本発明（請求項２３）は、検索条件入力手段３０において、利用者が例示した文書を検索キーとして入力する手段を含み、特徴情報抽出手段５０において、与えられた複合メディア文書から検索キーとして例示された複合メディア文書を抽出する入力文書特徴情報抽出手段と、利用者から例示された文書から検索情報の特徴情報を抽出する検索特徴情報抽出手段を含み、文書比較手段８０において、入力文書特徴情報抽出手段で抽出された入力文書特徴情報と、検索特徴情報抽出手段で抽出された検索特徴情報により、複合メディア文書と検索キーとして例示された複合メディア文書間の類似度を計算する類似度計算手段を含む。

【００２６】本発明（請求項２４）は、類似度計算手段において、検索条件入力手段３０で検索キーとして例示された文書を構成しているテキスト情報、画像情報、音声情報及び構造情報と、蓄積手段に蓄積されている複合メディア文書を構成しているテキスト情報、画像情報、音声情報及び構造情報との構成要素毎の類似性判定結果に基づく評価値を、複合メディア文書全体の類似度として設定する類似度設定手段を含む。

【００２７】本発明（請求項２５）は、類似度設定手段において、構成要素毎の類似性判定結果として、例示さ

れた文書のテキスト情報と蓄積手段に蓄積されている文書のテキスト情報の類似性判定結果と、該例示された文書の画像情報と、該蓄積手段に蓄積されている文書の画像情報との類似性判定結果と、該例示された文書の音声情報と該蓄積手段に蓄積されている文書の音声情報との類似性判定結果と、該例示さた文書の構造情報と該蓄積手段に蓄積されている文書の構造情報との類似性判定結果を用いる。

【００２８】本発明（請求項２６）は、類似度設定手段において、文書に含まれるテキスト情報、画像情報や音声情報、及び構造情報の特徴情報毎に、類似度を計算し、該類似度に重みの値を掛け、線形和をとったものを、複合メディア文書全体としての類似度とする線形和算出手段を含む。本発明（請求項２７）は、文書比較手段において、同一文書中に同一メディアが複数存在する場合に、例示された文書中に含まれるテキスト、画像、音声を含むメディア毎の全ての検索キーについて、蓄積された文書中のメディアにおける該検索キーに対する全ての類似度を計算し、該類似度が最も高いものを検索キーに対する代表の類似度とする代表類似度決定手段を含む。

【００２９】本発明（請求項２８）は、類似度計算手段において、検索キーとして、例示された文書の構造情報と蓄積された文書の構造情報のそれぞれを、順序ラベル付木（ordered labeled tree）として表現し、例示された文書を表現した順序ラベル付木と、蓄積された文書を表現した順序ラベル付木との形状を比較することで、文書の構造情報の類似度を設定する順序ラベル付木形状比較手段を含む。

【００３０】本発明（請求項２９）は、順序ラベル付木形状比較手段において、構造情報の類似度を判定する際に、文書構造を木と見做し、例示された文書を表す木から蓄積された文書を表す木へ変換するために必要なノードの挿入、ノードの削除、ノード名の変更を含む編集操作を行った回数と、該編集操作を行うのに必要なコストから算出する編集距離を文書の類似度として設定する編集距離算出手段を含む。

【００３１】本発明（請求項３０）は、類似度計算手段において、文書に含まれるテキスト情報、画像情報や音声情報、構造情報の特徴情報に基づく類似度を計算する手段と、構造情報の特徴情報に基づく類似度計算結果に基づく第１段階目の選択を行う手段と、文書に含まれるテキスト情報、画像情報や音声情報の特徴情報に基づく類似度を、複合メディア文書全体としての類似度とする手段とを含む。

【００３２】本発明（請求項３１）は、類似度計算手段において、文書に含まれるテキスト情報、画像情報や音声情報、構造情報の特徴情報に基づく類似度を計算する手段と、テキスト情報、画像情報や音声情報の特徴情報に基づく類似度計算結果に基づく第１段階目の選択を行う手段と、文書に含まれる構造情報の特徴情報に基づく類似度を、複合メディア文書全体としての類似度とする手段とを含む。

【００３３】本発明（請求項３２）は、類似度計算手段において、検索キーとして例示された文書中に同一メディアが複数存在する場合に、該メディアの文書レベルでの類似度を設定する文書レベル類似度計算手段を含む。本発明（請求項３３）は、文書レベル類似度計算手段において、検索キーとして例示された文書中に複数存在するメディアの各検索キーについて、例示された文書中に含まれるテキスト情報、画像情報、音声情報を含むメディア毎の全ての検索キーについて、蓄積された文書中の該メディアにおける該検索キーに対する全ての類似度を計算する手段と、類似度が最も高いものを検索キーに対する代表の類似度とする手段と、代表の類似度の平均値を計算する手段と、メディアの文書レベルでの類似度を設定する文書レベル類似度設定手段とを含む。

【００３４】本発明（請求項３４）は、文書レベル類似度設定手段において、検索キーとして例示された文書中に複数存在するメディアの各検索キーについて、該例示された文書中に含まれるテキスト情報、画像情報、音声情報を含むメディア毎の全ての検索キーについて、蓄積された文書中の該メディアにおける該検索キーに対する全ての類似度を計算する手段と、類似度が最も高いものを検索キーに対する代表の類似度とする手段と、代表の類似度のうち、最も類似度が高いものをメディアの文書レベルでの類似度とする手段とを含む。

【００３５】本発明（請求項３５）は、類似度計算手段において、検索キーとして例示された文書の構造情報と蓄積された文書の構造情報のそれぞれを順序ラベル付き木（ordered labeled tree）として表現し、それぞれの文書中の各メディアの特徴情報を順序ラベル付き木におけるノードの属性として格納した属性付き順序ラベル付き木として表現し、例示された文書を表現した属性付き順序ラベル付き木と蓄積された文書を表現した属性付き順序ラベル付き木との属性と形状を比較することで、複合メディア文書の類似度を設定する類似検索手段を含む。

【００３６】本発明（請求項３６）は、類似検索手段において、例示された文書を表現した属性付き順序ラベル付き木の各ノードの属性である特徴情報と類似した特徴情報を属性として持つノードを持つ蓄積された文書を表現した属性付き順序ラベル付き木について、該ノードの構造的な位置関係の差異から複合メディア文書の類似度を設定する手段を含む。

【００３７】本発明（請求項３７）は、順序ラベル付き木形状比較手段において、文書構造を順序ラベル付き木と見做し、該順序ラベル付き木に関する特徴情報に基づいて、多次元ベクトル空間上に該特徴情報を数値化してマッピングする手段と、ベクトル空間上での距離を文書

の類似度として設定する手段とを含む。

【００３８】本発明（請求項３８）は、順序付きラベル木に関する特徴情報として、順序ラベル付き木の各ノードの名前やノード数や各ノードの位置情報を数値化して利用することで文書の類似度を計算する。本発明（請求項３９）は、特徴情報として、テキスト情報であれば、テキストの記述内容が表す概念や各単語の出現頻度、画像情報であれば、画像の色相や彩度や輝度、色配置、音声情報であれば、音の強弱やメロディ、構造情報であれば、文書構造を順序ラベル付き木で表現した場合の木の形状やノードのラベル名、リンク情報などを、複合メディア文書の構成要素から抽出される特徴情報とする。

【００３９】本発明（請求項４０）は、文書比較手段８０において、例示された文書に対する蓄積された文書の類似度を、検索キーとして例示された文書を構成しているテキスト、画像、音声を含むメディアの情報及び構造情報と、蓄積されている文書を構成しているテキスト、画像、音声を含むメディアの情報及び構造情報との構成要素毎の類似性判定結果に基づく評価値を、複合メディア文書全体類似度として設定する手段と、蓄積された文書の類似度を降順に並べることで順位付けし、類似度を判定する手段とを含む。

【００４０】本発明（請求項４１）は、類似度設定手段において、複合メディア文書の各構成要素毎に、類似度を設定する手段と、文書に含まれるテキスト情報、画像情報や音声情報、及び構造情報の特徴情報毎に、類似度を計算する手段と、類似度に重みの値を掛け、線形和をとったものを、複合メディア文書全体としての類似度とする手段とを含む。

【００４１】本発明（請求項４２）は、類似度設定手段において、文書の構成要素毎の類似性判定結果に基づく評価値として、各構成要素の類似度そのもの、または、各構成要素の類似度に利用者から与えられた重みを掛け合わせたものを利用する手段を含む。本発明（請求項４３）は、テキスト情報、画像情報、人間の声のデータである音声データに加え、ＣＤやレコードを含む楽曲データや音楽データを含む音声情報から構成される構造化文書である複合メディア文書の類似検索プログラムを格納した記憶媒体であって、与えられた複合メディア文書及び、入力された検索条件を構文解析する文書解析プロセスと、文書解析プロセスで解析された結果、得られる文書の構成要素から特徴情報を抽出する特徴情報抽出プロセスと、複合メディア文書及び特徴情報抽出プロセスで抽出された特徴情報を記憶手段に格納する格納プロセスと、複合メディア文書を検索条件として入力させる検索条件入力プロセスと、記憶手段に蓄積されている複合メディア文書の特徴情報と、入力された検索条件を文書解析プロセスで解析した結果に基づいて特徴情報抽出プロセスで抽出された特徴情報から２つの複合メディア文書の類似度を判定する文書比較プロセスと、文書比較プロセスで判定された類似度に基づいて、類似した複合メディア文書を出力させる出力プロセスとを有する。

【００４２】本発明（請求項４４）は、検索条件入力プロセスにおいて、利用者が例示した文書を検索キーとして入力するプロセスを含み、特徴情報抽出プロセスにおいて、与えられた複合メディア文書から特徴情報を抽出する入力文書特徴情報抽出プロセスと、利用者から例示された文書から検索キーとして例示された複合メディア文書の特徴情報を抽出する検索特徴情報抽出プロセスを含み、文書比較プロセスにおいて、入力文書特徴情報抽出プロセスで抽出された入力文書特徴情報と、検索特徴情報抽出プロセスで抽出された検索特徴情報により、複合メディア文書と検索キーとして例示された複合メディア文書間の類似度を計算する類似度計算プロセスを含む。

【００４３】本発明（請求項４５）は、類似度計算プロセスにおいて、検索条件入力プロセスで検索キーとして例示された文書を構成しているテキスト情報、画像情報、音声情報及び構造情報と、記憶手段に蓄積されている複合メディア文書を構成しているテキスト情報、画像情報、音声情報及び構造情報との構成要素毎の類似性判定結果に基づく評価値を、複合メディア文書全体の類似度として設定する類似度設定プロセスを含む。

【００４４】本発明（請求項４６）は、類似度設定プロセスにおいて、構成要素毎の類似性判定結果として、例示された文書のテキスト情報と記憶手段に蓄積されている文書のテキスト情報の類似性判定結果と、該例示された文書の画像情報と、該記憶手段に蓄積されている文書の画像情報との類似性判定結果と、該例示された文書の音声情報と該記憶手段に蓄積されている文書の音声情報との類似性判定結果と、該例示さた文書の構造情報と該記憶手段に蓄積されている文書の構造情報との類似性判定結果を用いる。

【００４５】本発明（請求項４７）は、類似度設定プロセスにおいて、文書に含まれるテキスト情報、画像情報や音声情報、及び構造情報の特徴情報毎に、類似度を計算し、該類似度に重みの値を掛け、線形和をとったものを、複合メディア文書全体としての類似度とする線形和算出プロセスを含む。本発明（請求項４８）は、文書比較プロセスにおいて、同一文書中に同一メディアが複数存在する場合に、例示された文書中に含まれるテキスト、画像、音声を含むメディア毎の全ての検索キーについて、蓄積された文書中のメディアにおける該検索キーに対する全ての類似度を計算し、該類似度が最も高いものを検索キーに対する代表の類似度とする代表類似度決定プロセスを含む。

【００４６】本発明（請求項４９）は、類似度計算プロセスにおいて、検索キーとして、例示された文書の構造情報と蓄積された文書の構造情報のそれぞれを、順序ラベル付木（ordered labeled tree）として表現し、例示

された文書を表現した順序ラベル付木と、蓄積された文書を表現した順序ラベル付木との形状を比較することで、文書の構造情報の類似度を設定する順序ラベル付木形状比較プロセスを含む。

【００４７】本発明（請求項５０）は、順序ラベル付木形状比較プロセスにおいて、構造情報の類似度を判定する際に、文書構造を木と見做し、例示された文書を表す木から蓄積された文書を表す木へ変換するために必要なノードの挿入、ノードの削除、ノード名の変更を含む編集操作を行った回数と、該編集操作を行うのに必要なコストから算出する編集距離を文書の類似度として設定する編集距離算出プロセスを含む。

【００４８】本発明（請求項５１）は、類似度計算プロセスにおいて、文書に含まれるテキスト情報、画像情報や音声情報、構造情報の特徴情報に基づく類似度を計算するプロセスと、構造情報の特徴情報に基づく類似度計算結果に基づく第１段階目の選択を行うプロセスと、文書に含まれるテキスト情報、画像情報や音声情報の特徴情報に基づく類似度を、複合メディア文書全体としての類似度とするプロセスとを含む。

【００４９】本発明（請求項５２）は、類似度計算プロセスにおいて、文書に含まれるテキスト情報、画像情報や音声情報、構造情報の特徴情報に基づく類似度を計算するプロセスと、テキスト情報、画像情報や音声情報の特徴情報に基づく類似度計算結果に基づく第１段階目の選択を行うプロセスと、文書に含まれる構造情報の特徴情報に基づく類似度を、複合メディア文書全体としての類似度とするプロセスとを含む。

【００５０】本発明（請求項５３）は、類似度計算プロセスにおいて、検索キーとして例示された文書中に同一メディアが複数存在する場合に、該メディアの文書レベルでの類似度を設定する文書レベル類似度計算プロセスを含む。本発明（請求項５４）は、文書レベル類似度計算プロセスにおいて、検索キーとして例示された文書中に複数存在するメディアの各検索キーについて、例示された文書中に含まれるテキスト情報、画像情報、音声情報を含むメディア毎の全ての検索キーについて、蓄積された文書中の該メディアにおける該検索キーに対する全ての類似度を計算するプロセスと、類似度が最も高いものを検索キーに対する代表の類似度とするプロセスと、代表の類似度の平均値を計算するプロセスと、メディアの文書レベルでの類似度を設定する文書レベル類似度設定プロセスとを含む。

【００５１】本発明（請求項５５）は、文書レベル類似度設定プロセスにおいて、検索キーとして例示された文書中に複数存在するメディアの各検索キーについて、該例示された文書中に含まれるテキスト情報、画像情報、音声情報を含むメディア毎の全ての検索キーについて、蓄積された文書中の該メディアにおける該検索キーに対する全ての類似度を計算するプロセスと、類似度が最も高いものを検索キーに対する代表の類似度とするプロセスと、代表の類似度のうち、最も類似度が高いものをメディア文書中レベルでの類似度とするプロセスとを含む。

【００５２】本発明（請求項５６）は、類似度計算プロセスにおいて、検索キーとして例示された文書の構造情報と蓄積された文書の構造情報のそれぞれを順序ラベル付き木（ordered labeled tree）として表現し、それぞれの文書中の各メディアの特徴情報を順序ラベル付き木におけるノードの属性として格納した属性付き順序ラベル付き木として表現し、例示された文書を表現した属性付き順序ラベル付き木と蓄積された文書を表現した属性付き順序ラベル付き木との属性と形状を比較することで、複合メディア文書の類似度を設定する類似検索プロセスを含む。

【００５３】本発明（請求項５７）は、類似検索プロセスにおいて、例示された文書を表現した属性付き順序ラベル付き木の各ノードの属性である特徴情報と類似した特徴情報を属性として持つノードを持つ蓄積された文書を表現した属性付き順序ラベル付き木について、該ノードの構造的な位置関係の差異から複合メディア文書の類似度を設定するプロセスを含む。

【００５４】本発明（請求項５８）は、順序ラベル付き木形状比較プロセスにおいて、文書構造を順序ラベル付き木と見做し、該順序ラベル付き木に関する特徴情報に基づいて、多次元ベクトル空間上に該特徴情報を数値化してマッピングするプロセスと、ベクトル空間上での距離を文書の類似度として設定するプロセスとを含む。

【００５５】本発明（請求項５９）は、順序付きラベル木に関する特徴情報として、順序ラベル付き木の各ノードの名前やノード数や各ノードの位置情報を数値化して利用することで文書の類似度を計算する。本発明（請求項６０）は、特徴情報として、テキスト情報であれば、テキストの記述内容が表す概念や各単語の出現頻度、画像情報であれば、画像の色相や彩度や輝度、色配置、音声情報であれば、音の強弱やメロディ、構造情報であれば、文書構造を順序ラベル付き木で表現した場合の木の形状やノードのラベル名、リンク情報などを、複合メディア文書の構成要素から抽出される特徴情報とする。

【００５６】本発明（請求項６１）は、文書比較プロセスにおいて、例示された文書に対する蓄積された文書の類似度を、検索キーとして例示された文書を構成しているテキスト、画像、音声を含むメディアの情報及び構造情報と、蓄積されている文書を構成しているテキスト、画像、音声を含むメディアの情報及び構造情報との構成要素毎の類似性判定結果に基づく評価値を、複合メディア文書全体類似度として設定するプロセスと、蓄積された文書の類似度を降順に並べることで順位付けし、類似度を判定するプロセスとを含む。

【００５７】本発明（請求項６２）は、類似度設定プロ

セスにおいて、複合メディア文書の各構成要素毎に、類似度を設定するプロセスと、文書に含まれるテキスト情報、画像情報や音声情報、及び構造情報の特徴情報毎に、類似度を計算するプロセスと、類似度に重みの値を掛け、線形和をとったものを、複合メディア文書全体としての類似度とするプロセスとを含む。

【００５８】本発明（請求項６３）は、類似度設定プロセスにおいて、文書の構成要素毎の類似性判定結果に基づく評価値として、各構成要素の類似度そのもの、または、各構成要素の類似度に利用者から与えられた重みを掛け合わせたものを利用するプロセスを含む。上記のように、本発明では、与えられた文書の構文解析を行うことで、例示された文書と蓄積された文書の間で比較を行うべきメディアの情報及び構造情報などの構成要素の単位を決定することが可能となる。

【００５９】また、抽出した特徴情報により、文書を特徴付けることで、文書の内容や論理構造の情報に基づいた検索を可能にする。また、テキスト情報だけでなく、画像情報や音声情報、構造情報なども検索キーの一部として利用することが可能となる。さらに、蓄積されているテキスト情報、画像情報、音声情報、構造情報などから、それらを含んでいた文書への索引を作成することで、文書に高速にアクセスすることが可能となる。

【００６０】また、ディスプレイ上で文書のテキスト情報、画像情報などの特徴情報の内容を確認できるので、利用者が意図した特徴情報を含む複合メディア文書を検索キーとして入力することが容易である。また、検索キーとして入力された複合メディア文書に含まれる特徴情報毎に、類似度を計算し、それらに基づく評価値を計算する。例えば、文書の特徴情報毎に、類似度を計算し、それらに重みの値を掛け、足しあわせたものを文書レベルでの類似度として計算することで、テキスト情報以外に画像や音声情報、構造情報の類似性も検索条件として利用した複合メディア文書の類似検索が可能になる。

【００６１】また、文書中の特徴情報毎に、類似度を計算するので、個々の類似度計算方法に関して、例えば、画像情報の類似度計算方法だけ異なる類似度計算方法を採用し、部分的に置き換えるということが容易に行うことができる。

【００６２】

【発明の実施の形態】複合メディア文書の構成要素としては、図３に示すように、テキスト情報、画像情報、音声情報及び構造情報等がある。以下、当該複合メディア文書における類似検索について説明する。図４は、本発明の複合メディア文書の類似検索装置の構成を示す。

【００６３】同図に示す複合メディア文書の類似検索装置は、複合メディア文書入力装置１０、検索条件入力装置２０、検索条件入力部３０、複合メディア文書解析部４０、特徴情報抽出部５０、蓄積部６０、メモリ７０、文書比較部８０、表示装置９０から構成される。複合メディア文書入力装置１０は、テキスト情報、画像情報、音声情報及び構造情報を含む文書を入力する。

【００６４】検索条件入力装置２０は、利用者が入力のために利用するマウス等のポインティングデバイスや、キーボード等である。検索条件入力部３０は、利用者に検索条件入力装置２０であるキーボードから文書のファイル名を入力させたり、マウスを操作させて文書のアイコンをクリックさせたり、前回の検索結果で得られた文書をマウスでクリックさせることで検索キーとして入力する複合メディア文書を取得する。詳しくは、複合メディア文書を検索するための検索キーとなる複合メディア文書を例示する。例示する文書のファイル名を指定したり、例示する文書のアイコンをポインティングデバイスなどによりディスプレイ上でクリックすることで検索キーを例示する。また、文書を例示する際に、利用者が類似性を重視したい部分を指定することが可能であり、類似性を重視したい部分の特徴情報に対し、重視する度合いを示す重みの値を適宜変更して入力することが可能である。この時、文書中のどの部分の類似性を重視するかという重みの値と検索結果として返却する文書数ｋを利用者から取得する。あるいは、システムのデフォルト値を利用する。

【００６５】複合メディア文書解析部４０は、複合メディア文書入力装置１０または、検索条件入力部３０から与えられた文書の構文解析を行い、テキスト情報、画像情報、音声情報、構造情報などの文書の構成要素を検出する。複合メディア文書解析部４０は、ここで、ＳＧＭＬやＸＭＬのパーサ（parser：構文解析プログラム）を用いて入力された文書を解析し、文書からテキスト情報、画像情報、音声情報、構造情報等の文書の構成要素を検出する。

【００６６】特徴情報抽出部５０は、テキスト情報、画像情報、音声情報、構造情報などの文書の構成要素の特徴を表現している特徴情報を抽出する。例えば、テキスト情報ならテキストの記述内容が表す概念など、画像情報なら、画像情報の色相や彩度や輝度、色配置など、音声情報なら音の強弱やメロディなどの特徴情報を、特徴情報が格納されていた文書のＩＤ、エレメント名や出現位置の情報と共に抽出する。また、構造情報なら、文書構造を順序ラベル付き木で表現した場合の木の形状（階層構造など）やノードのラベル名、また、リンク情報などを複合メディア文書の構成要素から抽出される特徴情報とする。

【００６７】蓄積部６０は、与えられた文書をメモリ７０に蓄積する。また、各特徴情報から当該特徴情報を含んでいた文書への索引を作成する。文書比較部８０は、例示された複合メディア文書とメモリ７０に蓄積された複合メディア文書との特徴情報を比較することにより、類似度を求め、類似度の高いものを出力する。複合メディア文書としての類似度は、テキスト情報、画像情報、

音声情報、構造情報などの個々の類似度計算結果に基づいた評価値を計算したものとする。例えば、テキスト情報、画像情報、音声情報などの類似度に関しては、多次元ベクトル空間モデルに基づき、各特徴情報を多次元ベクトル空間上へマッピングし、多次元ベクトル空間上の例示された文書の特徴情報と蓄積された文書の特徴情報との２点間の距離が近ければ、類似度が高くなるように設定するというアプローチを採用することが可能である。また、蓄積された文書の類似度を降順に並べることで順位付けをし、類似度を判定することも可能である。

【００６８】以下、上記の構成における動作を複合メディア文書蓄積フェーズと、複合メディア文書検索フェーズに分けて説明する。図５は、本発明の複合メディア文書蓄積フェーズのフローチャートである。

ステップ１０１）　まず、複合メディア文書入力装置１０から複合メディア文書を入力する。

【００６９】ステップ１０２）　複合メディア文書解析部４０が、複合メディア文書入力装置１０から入力された複合メディア文書の構文解析を行い、テキスト情報、画像情報、音声情報、構造情報などの文書の構成要素を検出する。

ステップ１０３）　次に、特徴情報抽出部５０は、テキスト情報、画像情報、音声情報、構造情報などの文書構成要素について、例えば、テキスト情報なら、テキストの記述内容が表す概念など、画像情報なら画像の色相や彩度や輝度や色配置など、音声情報なら音の強弱やメロディなどの特徴情報を、特徴情報が格納されていた文書のＩＤ、エレメント名や出現位置の情報と共に抽出する。当該処理をすべての構成要素の数分繰り返す。

【００７０】ステップ１０４）　蓄積部６０は、与えられた文書及び、各特徴情報から当該特徴情報を含んでいた文書への索引を作成し、メモリ７０に格納する。次に、複合メディア文書検索フェーズの動作を説明する。図６は、本発明の複合メディア文書検索フェーズのフローチャートである。

ステップ２０１）　検索条件入力部３０は、検索条件入力装置２０であるキーボードから文書のファイル名を入力させたり、マウスを操作させて文書のアイコンをクリックさせたり、前回の検索結果で得られた文書をマウスでクリックさせることで、検索キーとして入力する複合メディア文書を取得する。この時、文書中のどの部分の類似性を重視するかという重みの値と、検索結果として返却する文書数ｋを利用者から取得する。あるいは、システムのデフォルト値を利用する。

【００７１】ステップ２０２）　次に、複合メディア文書解析部４０は、複合メディア文書蓄積フェーズの処理と同様に、複合メディア検索条件入力部３０から入力された複合メディア文書の構文解析を行い、テキスト情報、画像情報、音声情報、構造情報などの文書の構成要素を検出する。

ステップ２０３）　特徴情報抽出部５０が、複合メディア文書蓄積フェーズと同様に、テキスト情報、画像情報、音声情報、構造情報などの文書構成要素の特徴情報を、特徴情報が格納されていた文書のＩＤ、エレメント名や出現位置の情報と共に抽出し、例示された文書のテキスト情報、画像情報、音声情報、構造情報などの文書の構成要素について特徴情報を抽出する。当該処理をすべての構成要素の数分繰り返す。

【００７２】ステップ２０４）　文書比較部８０は、例示された文書の特徴情報とメモリ７０に蓄積された文書の特徴情報とを比較し、個々の特徴情報毎に類似度を計算し、それらの計算結果に基づいた評価値を複合メディア文書としての類似度として計算する。類似度の計算方法は後述する。

ステップ２０５）　文書比較部８０は、類似度を降順に並べ、利用者が要求した上位ｋ件の文書を類似度の高い文書として索引から選ぶ。

【００７３】ステップ２０６）　選択された類似度の高い文書を検索結果として表示装置９０に表示する。次に、上記における類似度を求める方法について説明する。図７は、本発明の類似度を求めるための文書比較を行う際のフローチャート（その１）である。

【００７４】ステップ３０１）　文書比較部８０は、特徴情報抽出部５０から検索条件入力部３０から入力された文書（検索条件）の特徴情報と、蓄積部６０から入力された複合メディア文書の特徴情報を取得する。

ステップ３０２）　特徴情報が構造情報である場合にはステップ３０３に移行し、そうでない場合にはステップ３０４に移行する。

【００７５】ステップ３０３）　文書比較部８０は、検索条件の特徴情報と複合メディア文書の特徴情報の構造情報を木と見做して、当該２つの木の間の編集距離を計算し、ステップ３０６に移行する。また、ノード間の構造的位置関係を計算する、木を多次元ベクトル化し、多次元ベクトル空間上の距離計算する等の方法も可能である。

【００７６】ステップ３０４）　文書比較部８０は、多次元ベクトル空間上の距離を計算する。

ステップ３０５）　同種の特徴情報のうち、距離が最小のものを代表として選択する。

ステップ３０６）　文書レベルでの類似度を計算する。

【００７７】ステップ３０７）　類似度が高い文書を索引から選択する。

ステップ３０８）　選択された文書を表示装置９０に出力する。上記の類似度計算の一方法として、例えば、以下のようにして類似度を求めることが可能である。

（１）　第１の類似度計算方法：検索キーとして例示された文書を構成しているテキスト、画像、音声を含むメディアの情報及び構造情報と、メモリ７０に蓄積されている文書を構成しているテキスト、画像、音声を含むメ

ディアの情報及び構造情報との構成要素毎の類似性判定結果に基づく評価値を、複合メディア文書全体類似度として設定する。ここで、類似判定結果に基づく評価値とは、各構成要素の類似度そのもの、または、各構成要素の類似度に利用者から与えられた重みを掛け合わせたものなどを利用する。

【００７８】（２）　第２の類似度計算方法：テキスト情報の類似度は、入力されたテキスト情報の特徴情報とメモリ７０に蓄積されたテキスト情報の特徴情報との多次元ベクトル空間上での距離を求めることで計算する。画像情報の類似度は、入力された画像情報の特徴情報とメモリ７０に蓄積された画像情報の特徴情報との多次元ベクトル空間上での距離を求めることで計算する。

【００７９】音声情報の類似度は、入力された音声情報の特徴情報とメモリ７０に蓄積された音声情報の特徴情報との多次元ベクトル空間上での距離を求めることで計算する。上記の各々の情報において、多次元ベクトル空間上での距離が小さいものが、類似度が高いものとして計算される。

【００８０】（３）　第３の類似度計算方法：また、上記の（２）の方法に加えて、同一文書中に同一メディアが複数存在する場合は、類似度が最も高いものを代表の類似度と設定する。例えば、図８に示すように、画像情報を複数含んでいる文書などでは、検索キーとして例示された文書中に存在する画像情報について、蓄積された文書中に複数存在する画像情報との類似度を計算し、その中で類似度が最も高いものを代表の類似度として設定する。これを、検索キーとして例示された文書中に存在するすべての画像情報について行う。

【００８１】図８において、例示文書中の画像Ａと蓄積文書中の画像ａ，ｂ，ｃとのそれぞれの類似度を求め、類似度が最も高いもの（例えば、画像ａ）を例示文書中の画像Ａに対する蓄積文書中の類似画像とする。さらに、例示文書中の画像Ｂと蓄積文書中の画像ａ，ｂ，ｃとのそれぞれの類似度を求め、類似度が最も高いもの（例えば、画像ｃ）を例示文書中の画像Ｂに対する蓄積文書中の類似画像とする。

【００８２】（４）　第４の類似度計算方法：また、特徴情報のうちの構造情報は、文書構造を木と見做し、一方の木からもう一方の木へ変換するために必要な編集距離を計算し、編集距離が小さければ類似度が高くなるように設定する。編集距離は、木を変換する際に必要なノードの挿入、ノードの削除、ノード名の変更という編集操作を行った回数と、それらの編集操作を行うのに必要なコストから算出する。これにより、類似度を計算することが可能であり、編集距離が小さいものが類似度の高いものとして計算される。

【００８３】（５）　第５の類似度計算方法：テキスト情報、画像情報、音声情報、構造情報などの類似度をそれぞれ計算し、検索条件入力部３０で取得した文書中のどの部分の類似性を重視するかという重みの値、あるいは、システムのデフォルト値に基づいて、テキスト情報、画像情報、音声情報、構造情報などの類似度それぞれに与えられた個別の重みの値を掛け、線形和をとる。この線形和をとったものが、複合メディア文書としての類似度に相当する。

【００８４】（６）　第６の類似度計算方法：次に、文書比較部８０において、複合メディア文書の類似度を計算する際に、文書に含まれるテキスト情報、画像情報や音声情報、構造情報の特徴情報に基づく類似度を計算し、構造情報の特徴情報に基づく類似度計算結果に基づいて、第１段階目の選抜を行った後で、文書に含まれるテキスト情報、画像情報や音声情報の特徴情報に基づく類似度を、複合メディア文書全体としての類似度とする。

【００８５】以下にこの方法を詳細に説明する。図９は、本発明の類似度を求めるための文書比較を行う際のフローチャート（その２）である。

ステップ４０１）　特徴情報情報抽出部５０において、入力された検索条件に対する特徴情報が入力される。

【００８６】ステップ４０２）　テキスト情報の類似度は、入力されたテキスト情報の特徴情報とメモリ７０に蓄積されたテキスト情報の特徴情報との多次元ベクトル空間上での距離を求めたり、入力されたテキスト情報の特徴情報と蓄積されたテキスト情報の特徴情報との出現頻度などから得られる値の差を求めることで計算する。

【００８７】画像情報の類似度は、入力された画像情報の特徴情報とメモリ７０に蓄積された画像情報の特徴情報との多次元ベクトル空間上での距離を求めることで計算する。音声情報の類似度は、入力された音声情報の特徴情報とメモリ７０に蓄積された音声情報の特徴情報の多次元ベクトル空間上での距離を求めることで計算する。なお、多次元ベクトル空間上での距離が小さいものや、出現頻度などから得られる値の差の絶対値が小さいものが、類似度が高いものとして計算される。

【００８８】また、特徴情報のうち、構造情報は、文書情報を木と見做し、一方の木からもう一方の木へ変換するために必要な編集距離を計算することや、木の特徴情報を数値化して多次元ベクトル化して多次元ベクトル空間上での距離を求めることなどで類似度を計算することが可能である。編集距離が小さいものや、多次元ベクトル空間上での距離が小さいものが、類似度の高いものとして計算される。

【００８９】ステップ４０３）　テキスト情報、画像情報、音声情報、構造情報などの類似度をそれぞれ計算し、検索条件入力部３０で取得した文書中のどの部分の類似性を重視するかという重みの値、あるいは、システムのデフォルト値に基づき、構造情報の類似度に基づく第１段階目の選抜を行う。

ステップ４０４）　第１段階目の選抜を行った後に残っ

た文書のテキスト情報、画像情報、音声情報などの類似度が、複合メディア文書としての類似度に相当する。

【００９０】ステップ４０５）　類似度が高い文書を索引から選択する。

ステップ４０６）　選択された文書を表示装置９０に表示する。

（７）　第７の類似度計算方法；次に、文書比較部８０において、複合メディア文書の類似度を計算する際に、文書に含まれるテキスト情報、画像情報や音声情報、構造情報の特徴情報に基づく類似度を計算し、テキスト情報、画像情報や音声情報の特徴情報に基づく類似度計算結果に基づく第１段階目の選抜を行った後で、文書に含まれる構造情報の特徴情報に基づく類似度を、複合メディア文書全体としての類似度とする。

【００９１】以下にこの方法を詳細に説明する。図１０は、本発明の類似度を求めるための文書比較を行う際のフローチャート（その３）である。

ステップ５０１）　入力された検索条件の特徴情報が入力される。

ステップ５０２）　テキスト情報の類似度は、入力されたテキスト情報の特徴情報とメモリ７０に蓄積されたテキスト情報の特徴情報との多次元ベクトル空間上での距離を求めたり、入力されたテキスト情報の特徴情報と蓄積されたテキスト情報の特徴情報との出現頻度などから得られる値の差を求めることで計算する。

【００９２】画像情報の類似度は、入力された画像情報の特徴情報と蓄積された画像情報の特徴情報との多次元ベクトル空間上での距離を求めることで計算する。音声情報の類似度は、入力された音声情報の特徴情報と蓄積された音声情報の特徴情報との多次元ベクトル空間上での距離を求めることで計算する。多次元ベクトル空間上での距離が小さいものや、出現頻度などから得られる値の差の絶対値が小さいものが類似度が高いものとして計算される。

【００９３】また、特徴情報のうち構造情報は、文書構造を木と見做し、一方の木からもう一方の木へ変換するために必要な編集距離を計算することや、木の特徴情報を数値化して多次元ベクトル化して多次元ベクトル空間上での距離を求めることなどで類似度を計算することが可能である。編集距離が小さいものや、多次元ベクトル空間上での距離が小さいものが、類似度が高いものとして計算される。

【００９４】ステップ５０３）　テキスト情報、画像情報、音声情報、構造情報などの類似度をそれぞれ計算し、検索条件入力部３０で取得した文書中のどの部分の類似性を重視するかという重みの値、あるいは、システムのデフォルト値に基づき、テキスト情報、画像情報、音声情報などの類似度に基づく第１段階目の選抜を行う。

【００９５】ステップ５０４）　第１段階目の選抜を行った後に、残った文書の構造情報の類似度が複合メディア文書としての類似度に相当する。

ステップ５０５）　類似度の高い文書を索引から選択する。

ステップ５０６）　選択された文書を表示装置９０に出力する。

（８）　第８の類似度計算方法：複合メディア文書の類似度を計算する際に、検索キーとして例示された文書中に同一メディアが複数存在する場合に、当該メディアの文書レベルでの類似度を設定する。例えば、上記の（３）では、蓄積された検索対象の文書中に同一メディアが複数存在する場合について述べたが、（８）では、検索キーとなる文書中に異なる画像が３つ存在する場合、検索キーとなる文書における画像の類似度をどう設定するのかという点について説明する。

【００９６】複合メディア文書の類似度を計算する際に、検索キーとして例示された文書中に同一メディアが複数存在する場合に、当該メディアの文書レベルでの類似度を設定する２種類の例に基づいて、文書中に画像情報が存在する場合について述べる。

①　メディアの文書レベルでの類似度を設定する際に、検索キーとして例示された文書中に複数存在する当該メディアの各検索キーについて、蓄積された文書中のメディアにおける検索キーに対する全ての類似度を計算し、類似度が最も高いものを検索キーに対する代表の類似度とし、代表の類似度の平均値を計算してメディアの文書レベルでの類似度とする場合に、図８に示す、例示文書中の画像Ａに対する蓄積文書中の類似画像として画像ａを得る。さらに、例示文書中の画像Ｂに対する蓄積文書中の類似画像として画像ｃを得る。画像Ａと画像ａとの間の類似度と、画像Ｂと画像ｃとの間の類似度の平均値を計算し、その値を例示文書と蓄積文書との間の文書レベルでの画像情報の類似度と設定する。

【００９７】②　メディアの文書レベルでの類似度を設定する際に、検索キーとして例示された文書中に複数存在する当該メディアの各検索キーについて、蓄積された文書中のメディアにおける検索キーに対する全ての類似度を計算し、類似度が最も高いものを検索キーに対する代表の類似度とし、代表の類似度のうち、最も類似度が高いものをメディアの文書レベルでの類似度とする場合に、図８に示す、例示文書中の画像Ａに対する蓄積文書中の類似画像として画像ａを得る。さらに、例示文書中の画像Ｂに対する蓄積文書中の類似画像として画像ｃを得る。画像Ａと画像ａとの間の類似度と、画像Ｂと画像ｃとの間の類似度のうち、最も類似度が高いもの（例えば、画像Ａと画像ａとの間の類似度）を例示文書と蓄積文書との間の文書レベルでの画像情報の類似度として設定する。

【００９８】次に、複合メディア文書の類似検索を行う際に、検索キーとして例示された文書の構造情報と蓄積

された文書の構造情報のそれぞれを順序ラベル付き木（ordered labeled tree）として表現する例について説明する。図１１は、本発明の複合メディア文書を属性付き順序ラベル付き木として表現することを説明するための図である。

【００９９】複合メディア文書の類似検索を行う際に、それぞれの文書中の各メディアの特徴情報を当該順序ラベル付き木におけるノードの属性として格納した属性付き順序ラベル付き木（順序ラベル付き木を拡張した木）として表現し、例示された文書を表現した属性付き順序ラベル付き木と蓄積された文書を表現した属性付き順序ラベル付き木との属性と形状を比較することで、複合メディア文書の類似度を設定する。

【０１００】この複合メディア文書の類似度を設定する際に、順序付きラベル付き木に関する特徴情報に基づいて、多次元ベクトル空間上に、当該特徴情報を数値化してマッピングし、当該ベクトル空間上での距離を文書の類似度として設定する。なお、特徴情報の数値化は、各ノードの名前（ラベル名）やノード数、各ノードの位置情報を数値化するものとする。

【０１０１】上述のように、利用者が詳細な文書構造を知らなくても、テキスト情報以外に構造情報も利用した文書の検索が行うことができる。また、テキスト情報の他に画像や音声の情報や構造情報も検索キーの一部に含めて文書の類似検索を行うことができる。また、図３に示す検索条件入力部３０、複合メディア文書解析部４０、特徴情報抽出部５０、蓄積部６０、文書比較部８０をプログラムとして構築し、複合メディア文書の類似検索装置として利用されるコンピュータに接続されるディスク装置や、フロッピーディスクやＣＤ－ＲＯＭ等の可搬記憶媒体に格納しておき、本発明を実施する際にインストールすることにより容易に本発明を実現できる。

【０１０２】なお、本発明は、上記の実施例に限定されることなく、特許請求の範囲内で種々変更・応用が可能である。

【０１０３】

【発明の効果】上述のように、本発明によれば、利用者が詳細な文書構造を知らなくても、テキスト情報以外の構造情報も利用した文書の検索を行うことができる。また、テキスト情報の他に画像や音声の情報や構造情報も検索キーの一部に含めて文書の類似を検索を行うことができる。

【図面の簡単な説明】

【図１】本発明の原理を説明するための図である。

【図２】本発明の原理構成図である。

【図３】本発明の複合メディア文書を説明するための図である。

【図４】本発明の複合メディア文書の類似検索装置の構成図である。

【図５】本発明の複合メディア文書蓄積フェーズのフローチャートである。

【図６】本発明の複合メディア文書検索フェーズのフローチャートである。

【図７】本発明の類似度を求めるための文書比較を行う際のフローチャート（その１）である。

【図８】本発明の同一文書中に同一メディアが複数存在する場合における類似度設定の方法を説明するための図である。

【図９】本発明の類似度を求めるための文書比較を行う際のフローチャート（その２）である。

【図１０】本発明の類似度を求めるための文書比較を行う際のフローチャート（その３）である。

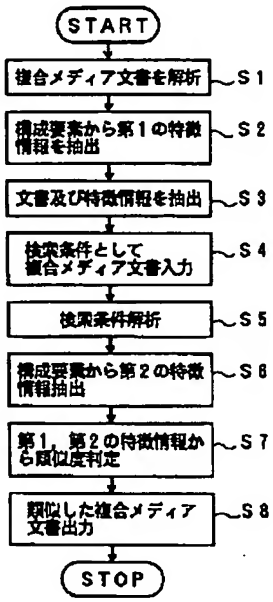
【図１１】本発明の複合メディア文書の属性付き順序ラベル付き木として表現することを説明するための図である。

【符号の説明】

１０　複合メディア文書入力手段、複合メディア文書入力装置
２０　検索条件入力装置
３０　検索条件入力手段、検索条件入力部
４０　文書解析手段、複合メディア文書解析部
５０　特徴抽出手段、特徴情報抽出部
６０　蓄積手段、蓄積部
７０　メモリ
８０　文書比較手段、文書比較部
９０　出力手段、表示装置

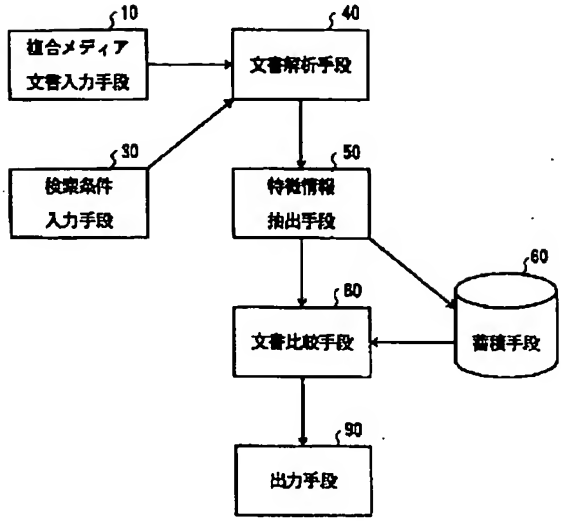
【図１】

本発明の原理を説明するための図



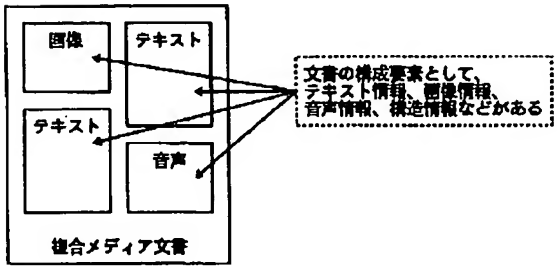

【図２】

本発明の原理構成図



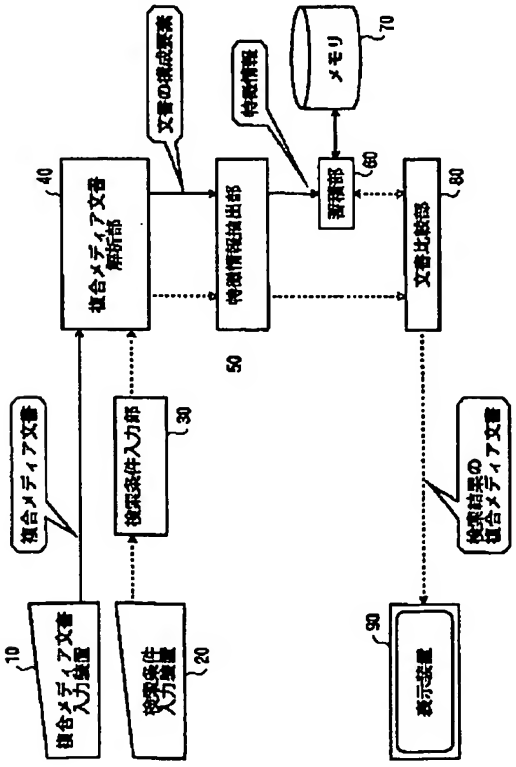

【図３】

本発明の複合メディア文書を説明するための図



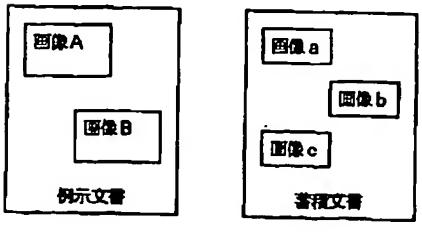

【図４】

本発明の複合メディア文書の類似検索装置の構成図



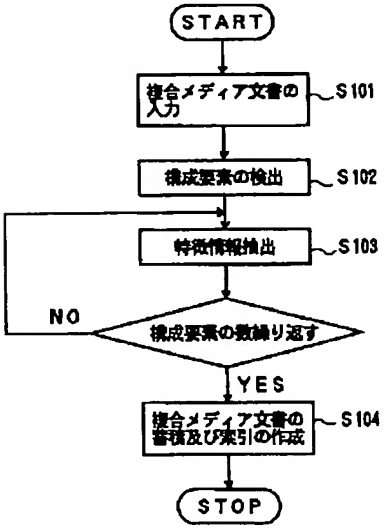

【図８】

本発明の同一文書中に同一メディアが複数存在する場合における類似度設定の方法を説明するための図



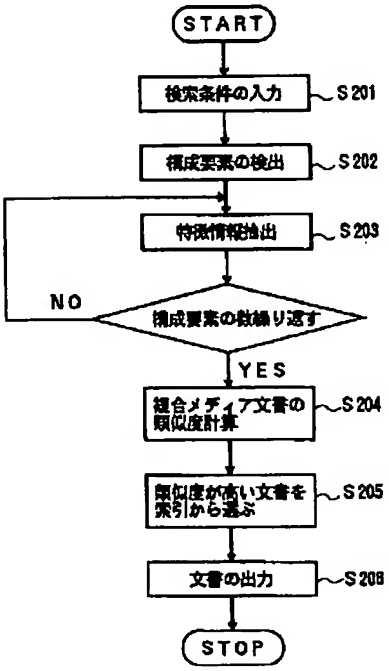
【図５】

本発明の複合メディア文書蓄積フェーズのフローチャート



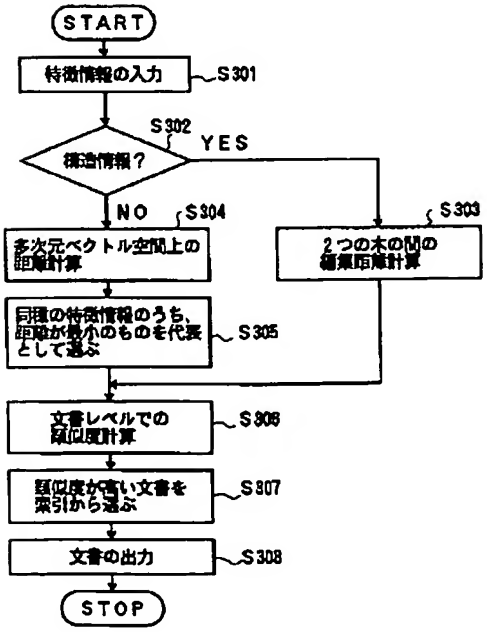

【図６】

本発明の複合メディア文書検索フェーズのフローチャート



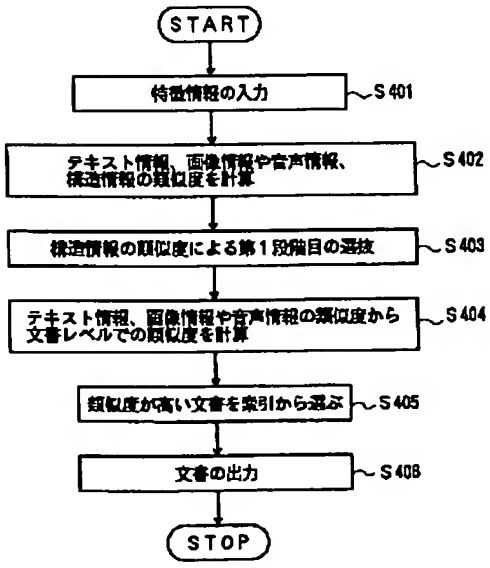

【図７】

本発明の類似度を求めるための文書比較を行う際のフローチャート（その１）



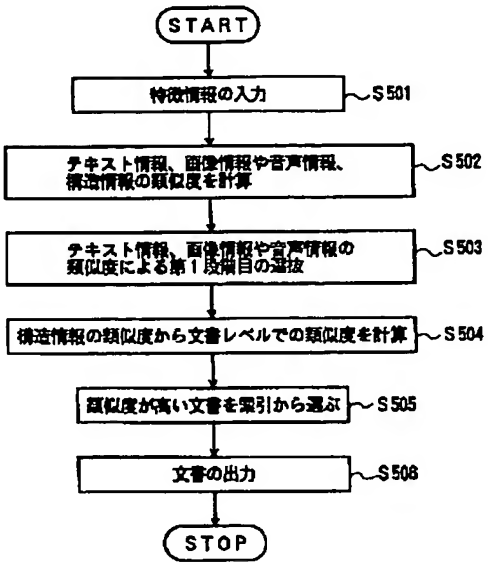

【図９】

本発明の類似度を求めるための文書比較を行う際のフローチャート（その２）



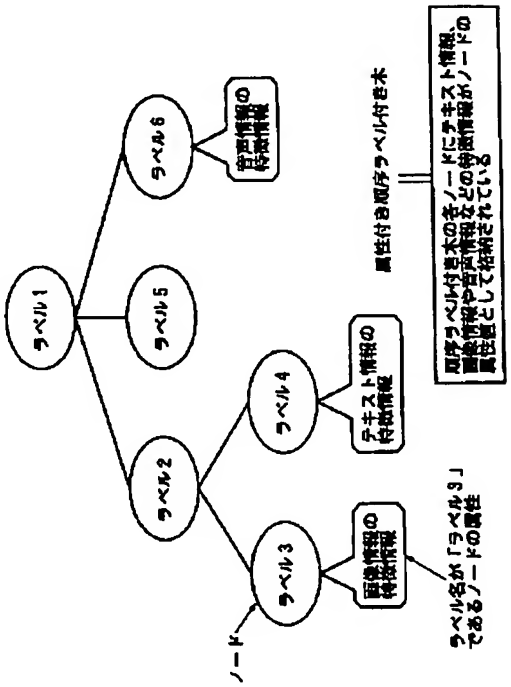
【図１０】

本発明の類似度を求めるための文書比較を行う際のフローチャート（その３）




【図１１】

本発明の複合メディア文書を属性付き順序ラベル付き木として表現することを説明するための図




フロントページの続き

(72)発明者　谷口　展郎
東京都新宿区西新宿三丁目19番２号　日本電信電話株式会社内

(72)発明者　山室　雅司
東京都新宿区西新宿三丁目19番２号　日本電信電話株式会社内

Ｆターム(参考)　5B075 ND03 ND06 ND14 ND16 PP24 PR06 QM08